A General Framework for Prediction in Generalized Additive Models

Alba Carballo González

Thesis submitted in partial compliance with the requirements for the degree of Doctor of Philosophy in Mathematical Engineering

Universidad Carlos III de Madrid

Advisors:

María Luz Durbán Reguera
Dae-Jin Lee Hwang

December 2019

A meu avó Agustín.

Acknowledgements

En primer lugar, quiero mostrar mi profunda gratitud a mis directores de tesis, María y Dae-Jin, porque el interés que pueda tener este documento se debe al tiempo y esfuerzo que habéis dedicado a su elaboración. María, sin tu paciencia ante mis errores esta tesis no habría sido posible, gracias por haberme guiado durante estos años y por haberme recibido con optimismo siempre que he tocado a tu puerta. Dae-Jin, gracias por tu ayuda y dedicación, por haberme hecho sentir comprendida (aunque eso implicase tener que corregir mis códigos), y por preocuparte por mí siempre que he estado en Bilbao. ¡Muchísimas gracias!

I want to express my gratitud to Göran and to the P-splines meeting members, Paul, Iain, Jutta, Giancarlo, Philipe and Gianluca, for your kindnees and your contributions, suggestions and comments on this work. Coté, moitísimas grazas por compartir conmigo os teus códigos e por resolverme tantas dúbidas.

Gracias al Departamento de Estadísitica y a sus profesores por instruirme en esta ciencia, a Gema, Susana y Paco por la ayuda siempre amable, y a Nacho por la organización e implicación en los Campus Científicos. Gracias a mis compañeros de máster y doctorado, Yoel, María, Antonio, Juan Carlos y José Antonio, por el apoyo y la ayuda, tanto con la docencia como con la investigación. Maicol, no podría haber tenido mejor compañero de despacho, siempre recordaré los buenos momentos en el 7.3.J10, ¡gracias! Eli, MJ, Rubén y Mañas, gracias por haber sido mi familia en Leganés durante dos años y por seguir contribuyendo a que mi capacidad de resiliencia sea mayor. Diego, gracias por haberme enseñado mucho más que P-splines, y por estar siempre dispuesto a ayudarme.

Quero agradecer aos profesores, estudantes e compañeiros que contribuiron a que a miña andaina académica estivera chea de bos momentos, e que se preocuparon por min durante estes anos. Especialmente a Elisa, porque contigo comecei a descubrir o xeniais que son as matemáticas, e a Javier, Pablo e Marco, por haberme transmitido vuestra curiosidad e ilusión.

Grazas ás persoas que tiven a sorte de atopar polo camiño, e que comparten o seu tempo conmigo facendo que todo sexa mellor, Aida, Alex, Cris, Enrique, María, Nati, Silvia e Yoli. O xeral do agradacemento non implica que descoñeza todo o que vos debo, algúns levades once anos facéndome sentir preto de vós a pesar da distancia, grazas de verdade. Hollis, thanks for sharing with me your concerns and achievements, and for your encouragement and help.

Para rematar, quero agradecer á miña familia. Non foi doado non poder estar con vós sempre que debería ou que me gustaría, pero serviume para valorar, máis se cabe, o afortunada que son por tervos. Grazas á tía María, e a Marta, Lourdes e Saleta pola axuda e as mostras de agarimo que teño recibido. Moitísimas grazas a miña avóa Mercedes, a meus tíos, Manolo, Ana, Pepe e Maruxa, e a meus curmáns Fiz, Iago e Lola, por preocuparvos por min, por darme os mellores consellos e por axudarme sempre que o necesito, para min as vosas mensaxes foron apertas. Madriña, grazas por ser o mellor exemplo e por coidarme tanto. Alberto, grazas por aguantarme como irmá maior e ao mesmo tempo facer que pase os mellores momentos, coidarme e axudarme a ser forte, quérote moito. Papás, grazas, entre outras moitas razóns, por escoitarme, entenderme e apoiarme, por ser a miña tranquilidade e por suplir as miñas carencias de loita e optimismo sempre que o necesito. Se nalgún momento estades orgullosos de Alberto e de min, estádeo moito máis de vós. Mòltes gràcies.

The research presented in this thesis has been partially supported by the Basque Government through the BERC 2018-2021 program and by Spanish Ministry of Economy and Competitiveness MINECO through BCAM Severo Ochoa excellence accreditation SEV-2013-0323 and through projects MTM2017-82379-R, funded by (AEI/FEDER, UE) and acronym “AFTERAM”, and MTM2014-52184-P.

Published contents

- A. Carballo, M. Durbán, D.-J. Lee. A general framework for prediction in penalized regression. UC3M Working Papers Statistics and Econometrics, 17-11, 2017. https://e-archivo.uc3m.es/handle/10016/24607
  – Co–author.
  – It is totally included in Chapter 2 of the thesis.
  – The material from this source included in the thesis is not indicated by typographical means or references.

- A. Carballo, M. Durbán, D.-J. Lee. Out-of-sample prediction in multidimensional P-spline models. UC3M Working Papers Statistics and Econometrics, 19-10, 2019. https://e-archivo.uc3m.es/handle/10016/28630
  – Co–author.
  – It is totally included in Chapters 3 and 4 of the thesis.
  – The material from this source included in the thesis is not indicated by typographical means or references.

- A. Carballo, M. Durbán, D.-J. Lee, P. Eilers. The memory of extrapolating P-splines. Procedings of the 31st International Workshop on Statistical Modelling, 2, 11-14, 2016. http://dupuy.perso.math.cnrs.fr/IWSM2016/proceedingsVol2.pdf
  – Co–author.
  – It is totally included in Chapters 2 of the thesis.
  – The material from this source included in the thesis is not indicated by typographical means or references.

Abstract

Smoothing techniques have become one of the most popular modelling approaches in the unidimensional and multidimensional setting. However, out-of-sample prediction in the context of smoothing models is still an open problem that can significantly widen the use of these models in many areas of knowledge. The objective of this thesis is to propose a general framework for prediction in penalized regression, particularly in the P-splines context.
To that end, Chapter 1 includes a review of the different proposals available in the literature, and results useful and necessary along the document, the formulation of a P-spline model and its reparameterization as a mixed model.

In Chapter 2, we generalize the approach given by Currie et al. (2004) to predict with any regression basis and quadratic penalty. For the particular case of penalties based on differences between adjacent coefficients, we reparameterize the extended P-spline model as a mixed model and we prove that the fit remains the same as the result we obtain only fitting the data and show the crucial role of the penalty order, since it determines the shape of the prediction. Moreover, we adapt available methods in contexts such as mixed models (Gilmour et al. 2004) or global optimization (Sacks et al. 1989) to predict in the context of penalized regression and prove their equivalence for the particular case of P-splines. An extensive section of examples illustrates the application of the methodology. We use three real datasets with particular characteristics: one of them on aboveground biomass allow us to show that prediction can also be performed to the left of the data; other of them, on monthly sulphur dioxide levels, illustrates how prediction can take into account the temporal trends and seasonal effects by using the smooth modulation model based on P-splines suggested by Eilers et al. (2008); and other, on yearly sea level, shows that prediction can also be carried out in the case of correlated errors. We also introduce the concept of "memory of a P-spline" as a tool to know how much of the known information we use to predict new values.

In the third chapter, we propose a general framework for prediction in multidimensional

smoothing, we extend the proposal of Currie et al. (2004) to predict when more than one covariate is extended. The extension of the prediction method to the multidimensional case is not straightforward in the sense that, in this context, the fit changes when the fit and the prediction are carried out simultaneously. To overcome this problem we propose an easy but elegant solution, based on Lagrange multipliers. The first part of the chapter is dedicated to show how out-of-sample predictions can be carried out in the context of multidimensional P-splines and the properties satisfied, under certain conditions, by the coefficients that determine the prediction. We also propose the use of restrictions to maintain the fit, and in general, to incorporate any known information about the prediction. The second part of the chapter is dedicated to extend the methodology to the smooth mixed model framework. It is known that when a P-spline model is reparameterized as a mixed model, the structure of the coefficients is lost, that is, they are not ordered according to the position of the knots. This fact is not relevant when we fit the data, but if we predict and impose restrictions over the coefficients, we need to differentiate between the coefficients that determine the fit and the coefficients that determine the prediction. In order to do that, we define a particular transformation matrix that preserves the original model matrices. The prediction method and the use of restrictions is illustrated with one real data example on log mortality rates of US male population. We show how to solve the crossover problem of adjacent ages when mortality tables are forecasted and compare the results with the well-known method developed in Delwarde et al. (2007).

The research in Chapter 4, is motivated by the need to extend the prediction methodology in the multidimensional case to more flexible models, the so-called Smooth-ANOVA models, which allow us to include interaction terms that can be decomposed as a sum of several smooth functions. The construction of these models through B-splines basis suffer from identifiability problems. There are several alternatives to solve this problem, here we follow Lee and Durbán (2011) and reparameterize them as mixed models. The first two sections of the chapter are dedicated to introduce the Smooth-ANOVA models and to show how out-of-sample prediction can be carried out in these models. We illustrate the prediction with Smooth-ANOVA models reanalyzing the dataset on aboveground biomass. Now, the Smooth-ANOVA model allows us to represent the smooth function as the sum of a smooth function for the height, a smooth function for the diameter of a tree, and a smooth term for the height-diameter interaction. At the end of this chapter, we provide a simulation study in order to evaluate the accuracy of the 2D interaction P-spline models and Smooth-ANOVA models, with and without imposing invariance of the fit. From the results of the simulation study, we conclude that in most situations the constrained S-ANOVA model behaves better in the fit and out-of-sample predictions,

however, results depend on the simulation scenario and on the number of dimensions in which the prediction is carried out (one or both dimensions).

In the fifth chapter we generalize the developed methodology for generalized linear models (GLMs) in the context of P-splines (P-GLMs) and mixed models (P-GLMMs). In both frameworks, the coefficients and parameters estimation procedures involve nonlinear equations. To solve them iterative algorithms based on the Newton-Raphson methods are used, regardless of the estimation criterion used (for instance, in the GLMMs context we can maximize the residual maximum likelihood (REML) or an approximate REML (based on Laplace approximation)). These iterative algorithms are based on a working normal theory model or a set of pseudodata and weights. Based on this idea, we extend the Penalized Quasilikelihood method (PQL) to fit and predict simultaneously in the context of GLMM. We highlight that, in the context of mixed models (even in the univariate case), to maintain the fit a transformation that preserves the original model matrices has to be used, since different transformations deal with different working vectors and therefore with different solutions. We also show how restrictions can be imposed in P-GLMs and P-GLMMs models. To illustrate the procedures we use a real dataset to predict deaths due to respiratory disease through 2D interaction P-splines and S-ANOVA models (both with and without the restriction the fit has to be maintained).

Finally, Chapter 6 is devoted to summarize the main conclusions and pose a list of future lines of work.

Resumen

Las técnicas de suavizado se han convertido en uno de los enfoques de modelado más populares en el entorno unidimensional y multidimensional. Sin embargo, la predicción fuera del rango de valores conocidos en el contexto de los modelos de suavizado sigue siendo un problema abierto que puede ampliar significativamente el uso de estos modelos en muchas áreas de conocimiento. El objetivo de este documento es proponer un marco general para la predicción en regresión penalizada, particularmente en el contexto de P-splines.
Con ese fin, el Capítulo 1 incluye una revisión de las diferentes propuestas disponibles en la literatura y los resultados útiles y necesarios a lo largo del documento, la formulación de un modelo P-spline y su reparametrización como modelo mixto.

En el Capítulo 2, generalizamos el enfoque dado por Currie et al. (2004) para predecir con cualquier base de regresión y penalización cuadrática. Para el caso particular de penalizaciones basadas en diferencias entre coeficientes adyacentes, reparametrizamos el modelo P-spline extendido como un modelo mixto y demostramos que el ajuste sigue siendo el mismo que el resultado que obtenemos solo ajustando los datos, también mostramos el papel crucial del orden de penalización, ya que determina la forma de la predicción. Además, adaptamos los métodos disponibles en contextos como modelos mixtos (Gilmour et al. 2004) u optimización global (Sacks et al. 1989) predecir en el contexto de la regresión penalizada y demostramos su equivalencia para el caso particular de P-splines. Una extensa sección de ejemplos ilustra la aplicación de la metodología. Utilizamos tres conjuntos de datos reales con características particulares: uno de ellos, sobre biomasa, nos permite mostrar que la predicción también se puede realizar a la izquierda de los datos; otro de ellos, sobre los niveles mensuales de dióxido de azufre, ilustra como la predicción puede tener en cuenta las tendencias temporales y los efectos estacionales utilizando el modelo de modulación suave basado en P-splines sugerido por Eilers et al. (2008); y otro, sobre el nivel anual del mar, muestra que la predicción también se puede realizar en el caso de errores correlacionados. También presentamos el concepto de "memoria de un P-spline" como una herramienta para saber cuánta información conocida usamos para

predecir nuevos valores.

En el tercer capítulo, proponemos un marco general para la predicción en el suavizado multidimensional, ampliamos la propuesta de Currie et al. (2004) para predecir cuando se extiende más de una covariable. La extensión del método de predicción al caso multidimensional no es directa en el sentido de que, en este contexto, el ajuste cambia cuando el ajuste y la predicción se llevan a cabo simultáneamente. Para resolver este problema, proponemos una solución fácil, basada en multiplicadores de Lagrange. La primera parte del capítulo está dedicada a mostrar como se pueden realizar predicciones fuera de la muestra en el contexto de P-splines multidimensionales y las propiedades que satisfacen, bajo ciertas condiciones, los coeficientes que determinan la predicción. También proponemos el uso de restricciones para mantener el ajuste y, en general, para incorporar cualquier información conocida sobre la predicción. La segunda parte del capítulo está dedicada a extender la metodología al marco de modelos mixtos suaves. Se sabe que cuando un modelo de P-spline se reparametriza como un modelo mixto, la estructura de los coeficientes se pierde, es decir, no se ordenan de acuerdo con la posición de los nodos. Este hecho no es relevante cuando ajustamos los datos, pero si predecimos e imponemos restricciones sobre los coeficientes, necesitamos diferenciar entre los coeficientes que determinan el ajuste y los coeficientes que determinan la predicción. Para hacer eso, definimos una matriz de transformación particular que conserva las matrices del modelo original. El método de predicción y el uso de restricciones se ilustran con un ejemplo de datos reales sobre el logaritmo de las tasas de mortalidad de la población masculina estadounidense. Mostramos como resolver el problema de cruce de proyecciones edades adyacentes cuando se predicen tablas de mortalidad y comparamos los resultados con el método desarrollado en Delwarde et al. (2007).

La investigación en el Capítulo 4 está motivada por la necesidad de extender la metodología de predicción en el caso multidimensional a modelos más flexibles, los modelos Smooth-ANOVA, que nos permiten incluir términos de interacción que pueden descomponerse como una suma de varias funciones suaves. La construcción de estos modelos a través de B-splines tiene problemas de identificabilidad. Hay varias alternativas para resolver este problema, nosotros seguimos Lee and Durbán (2011) y lo reparametrizamos como modelos mixtos. Las primeras dos secciones del capítulo están dedicadas a presentar los modelos Smooth-ANOVA y mostrar como se puede llevar a cabo la predicción fuera del rango de valores observados en estos modelos. Ilustramos la predicción con modelos Smooth-ANOVA reanalizando el conjunto de datos sobre biomasa. Ahora, el modelo Smooth-ANOVA nos permite representar la función suave como la suma de una función suave para la altura, un término suave para el diámetro y una función suave para la

interacción altura-diámetro. Al final de este capítulo, proporcionamos un estudio de simulación para evaluar la precisión de los modelos de interacción 2D P-spline y los modelos Smooth-ANOVA, con y sin imponer la invariancia del ajuste. A partir de los resultados del estudio de simulación, concluimos que en la mayoría de las situaciones el modelo S-ANOVA restringido se comporta mejor tanto en el ajuste como en la predicción, sin embargo, los resultados dependen del escenario de simulación y del número de dimensiones en las que se realiza la predicción (una o ambas dimensiones).

En el quinto capítulo generalizamos la metodología desarrollada para modelos lineales generalizados (GLM) en el contexto de P-splines (P-GLM) y modelos mixtos (P-GLMM). En ambos marcos, los procedimientos de estimación de coeficientes y parámetros involucran ecuaciones no lineales. Para resolverlos, se utilizan algoritmos iterativos basados en los métodos de Newton-Raphson, independientemente del criterio de estimación utilizado (por ejemplo, en el contexto de GLMMs podemos maximizar la máxima verosimilitud residual (REML) o un REML aproximado (basado en la aproximación de Laplace)). Estos algoritmos iterativos se basan en un modelo teórico normal o en un conjunto de pseudodatos y pesos. Basándonos en esta idea, ampliamos el método Penalized Quasi-likelihood (PQL) para ajustar y predecir simultáneamente en el contexto de GLMMs. Destacamos que, en el contexto de modelos mixtos (incluso en el caso univariante), para mantener el ajuste, se debe utilizar una transformación que conserve las matrices del modelo original, ya que las diferentes transformaciones tratan con diferentes vectores de trabajo y, por lo tanto, con diferentes soluciones. También mostramos como se pueden imponer restricciones en los modelos P-GLM y P-GLMM. Para ilustrar los procedimientos, utilizamos un conjunto de datos real para predecir las muertes por enfermedad respiratoria a través de modelos 2D P-splines y modelos S-ANOVA (con y sin la restricción el ajuste debe mantenerse).

Finalmente, el Capítulo 6 se dedica a resumir las principales conclusiones y a plantear una lista de futuras líneas de trabajo.

Contents

List of figures

Chapter 1

Introduction

Additive models are a class of non-parametric regression methods which have been found widespread applications in practice. This is due to their ability to represent non-linear associations between covariates and response variables in an intuitive way. One of the main assumptions of additive models is that the effect of covariates on the dependent variable follows an additive form and the separate effects are modelled by smoothing functions. Additive models can be used to model and predict in many areas such as Epidemiology, Agriculture, Demography or Engineering.
Moreover, out-of-sample prediction in the context of smoothing model is a problem that is still unresolved and that can significantly improve the use of these models in many areas of knowledge. This can have a major impact in areas such as Demography (mortality tables) or spatio-temporal modelling, in which case we can be interested in prediction for two covariates (latitude and longitude). These are some of the main reasons that encourage us to work in the prediction field.

There are several methods that allow us to obtain the smooth functions that describe the mean of the response variable as a function of the regressors, i.e. the functions that determine the additive model. Hastie and Tibshirani (1990) gives a broad overview on smoothers, such as running-means, locally-weighted running-lines, kernels, regression spline and smoothing splines. However, important tools such as kernel smoothers are not so commonly used, and within the framework of splines (regression spline and smoothing splines) there are two main drawbacks, in the smoothing splines, the number of parameters is the same as the number of observations, and in the regression splines we face the difficulty of choosing the number and position of the knots. To deal with the previous drawbacks we will focus on the smoothing approach introduced by Eilers and Marx (1996). It is called splines with penalties (commonly known as P-splines). In the P-spline methodology we do not assume a rigid form for the dependence of the independent variable on the regressors, however the smooth functions are determined

by parameters, that is the reason why the methodology is classified as semi-parametric regression.

Eilers and Marx (1996) have simplified the approach developed in O'Sullivan (1988) and propose a methodology that combines B-splines (the number of parameters is much less than the dimension of the data) and a penalty that penalizes the jumps between adjacent coefficients (the number and position of knots is not crucial).
Since this is the framework that we will use through the entire thesis, in Section 1.2 we give a brief introduction to the P-splines methodology and its reparameterization as mixed models. But first, we review the main literature related to out-of-sample prediction in smooth models.

1.1 Preliminaries on prediction in smooth models

There are many situations in which prediction of new observations in the context of regression is needed, in particular, when "out of range" prediction is required (beyond the range of observed covariates). This problem extends in the framework of smoothing, i.e. for models where the regression function is a smooth but otherwise unspecified function. Most of the existing literature in the area is related to the prediction of new observations in a temporal context, i.e. forecast of new observations. We start by providing a brief review of the main literature related to forecasting in smoothing models by commenting the main approaches of Sacks et al. (1989) and Ba et al. (2012). We also refer to Hyndman et al. (2008) who give an overview of exponential smoothing methods.
Exponential smoothing or weighted smoothing, respectively, refers to a class of forecasting methods, each of them having the property that forecasts are weighted combinations of past observations, where the weights decrease exponentially as observations come from further in the past. The weights are given through a smoothing parameter, $0 \leq \alpha \leq 1$, being the weight for the last observation α, for the penultimate observation $\alpha(1-\alpha)$, for the third from last $\alpha(1-\alpha)^2$ and so on. Hyndman et al. (2008) provide extensive information about exponential smoothing methods.
A similar prediction problem is tackled in the framework of global optimization where the interest is to evaluate an unknown function at point x, say. The question is now where to place future values of x to evaluate the function, such that relevant (preferably most) information about the function is achieved. As an example we refer to Sacks et al. (1989) and Jones et al. (1998), who fit a stochastic process to data and predict the process at a new point given the already observed data. They treat the observations as if they were generated by a constant and an error component which is modelled as a stochastic

process. Their approach is called *Bayesian global optimization* and the concept is the same as the idea behind the well-known technique in spatial statistics called kriging Cressie (1993).
Prediction can be addressed in a Bayesian setting using Bayesian P-splines by exploiting the properties of the random walk prior used for the coefficients (Besag et al. 1995; Lang and Brezger 2004). Computation of the predicted pattern and its credible intervals can be performed via MCMC, or giving "infinity" prior on the variance of the missing-data. A recent strategy is to use penalized splines to fit and forecast time series data (Ba et al. 2012). In this case, one minimizes the penalized least squares criteria:

$$S = (\boldsymbol{y} - \boldsymbol{B\theta})'\boldsymbol{M}(\boldsymbol{y} - \boldsymbol{B\theta}) + \lambda\boldsymbol{\theta}'\boldsymbol{P\theta},$$

where $\boldsymbol{y}$ is the vector of observed responses, $\boldsymbol{\theta}$ are basis coefficients, $\boldsymbol{B}$ is a spline basis that covers the hole range of the explanatory variable and $\boldsymbol{M}$ is a weight matrix that assigns exponentially decreasing weights on the samples, according to the order of their arrival. Matrix $\boldsymbol{P}$ is a penalty matrix controlling the smoothness of the fitted function and λ is the smoothing parameter. Ba et al. (2012) propose an adaptive learning algorithm that updates the smoothing functions of additive models and that can be used to compute predictions for new given covariates. Currie et al. (2004) have shown how the method of penalized splines (P-splines), introduced by Eilers and Marx (1996) and extensively discussed in Ruppert et al. (2003), can be extended to smooth and predict simultaneously two-dimensional mortality tables. In particular, the authors show how to construct the appropriate regression bases and penalty matrices for forecasting. Camarda (2012) has implemented the code that allow us to fit and forecast Poisson counts with P-splines, but it really does not perform the fit and the forecast simultaneously. First the fit is performed and then, using the smoothing parameter estimated for the fit and a new set of knots that covers the range of the extended covariate (which does not have to contain the subset of the knots used to fit the data) the fit and the forecast are obtained. Ugarte et al. (2012) and Etxeberria et al. (2015) have carried out prediction of future observations in time in the context of P-splines using Currie et al. (2004). In order to preserve the fit when the fit and the forecast are obtained simultaneously they proposed a modification of the penalty matrix.

The aforementioned papers describe most of the literature related to prediction. As we have said, we will base our work on the P-splines framework and therefore on the forecasting approach given by Currie et al. (2004); in the next section, we summarize the basic concepts related to P-splines as well as those related to its reparameterization as mixed models.

1.2 P-splines methodology

In this section, we introduce the approach of smoothing done by Eilers and Marx (1996) (P-splines) and its reparameterization as mixed models. The main ideas that hold the P-splines methodology are: use a low-rank regression basis and penalize the difference between adjacent coefficients to control the smoothness.

1.2.1 P-splines

Suppose that we estimate a smooth model from the observed data pairs (x_i, y_i), $i = 1, ..., n$, with $\boldsymbol{y}$ a Gaussian response variable and $\boldsymbol{x}$ the regressor, we have the model

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}), \tag{1.1}$$

with $\boldsymbol{R}$ the variance-covariance matrix of errors. The errors can be independent and identically distributed (i.i.d.), i.e. $\boldsymbol{R} = \sigma^2_{\boldsymbol{\epsilon}} \boldsymbol{I}$, but in order to propose a general approach we consider $\boldsymbol{R}$ as any variance-covariance matrix and, to simplify the notation, we consider $\tilde{\boldsymbol{R}} = \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}} \boldsymbol{R}$. The aim is to estimate the function f, that is assumed to be smooth.

Writing the model (1.1) in matrix form, we have:

$$\boldsymbol{y} = \boldsymbol{B\theta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{R}) \tag{1.2}$$

where $\boldsymbol{B}$ is a regression basis constructed from the regressor $\boldsymbol{x}$, and $\boldsymbol{\theta}$ is the vector of regression coefficients. To obtain the coefficients, $\boldsymbol{\theta}$, Eilers and Marx (1996) proposed to minimize the following penalized sum squares problem:

$$S_p(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{B\theta})' \tilde{\boldsymbol{R}}^{-1} (\boldsymbol{y} - \boldsymbol{B\theta}) + \lambda \boldsymbol{\theta}' \boldsymbol{P} \boldsymbol{\theta}, \tag{1.3}$$

where $\lambda \boldsymbol{P}$ is the penalty, with $\boldsymbol{P}$ a matrix that penalizes the difference between adjacent coefficients and λ a smoothing parameter that controls the amount of smoothness. The solution of the penalized sum of squares (1.3), is:

$$\frac{\partial S_p}{\partial \boldsymbol{\theta}} = -2\boldsymbol{B}' \tilde{\boldsymbol{R}}^{-1} (\boldsymbol{y} - \boldsymbol{B\theta}) + 2\lambda \boldsymbol{P\theta} \Rightarrow \hat{\boldsymbol{\theta}} = (\boldsymbol{B}' \tilde{\boldsymbol{R}}^{-1} \boldsymbol{B} + \lambda \boldsymbol{P})^{-1} \boldsymbol{B}' \tilde{\boldsymbol{R}}^{-1} \boldsymbol{y}.$$

Notice that the size of this system of equations depends on the size of the basis and not on the number of observations. Therefore, we have that:

$$\hat{\boldsymbol{y}} = \boldsymbol{B}\hat{\boldsymbol{\theta}} = \boldsymbol{B}(\boldsymbol{B}' \tilde{\boldsymbol{R}}^{-1} \boldsymbol{B} + \lambda \boldsymbol{P})^{-1} \boldsymbol{B}' \tilde{\boldsymbol{R}}^{-1} \boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}, \tag{1.4}$$

where $\boldsymbol{H}$ is called smoother matrix or hat matrix of the model, its trace corresponds to the effective dimension of the model, a measure of complexity of the model defined by Hastie and Tibshirani (1990).

Regression basis

There are several alternatives for the choice of the regression basis $\boldsymbol{B}$ in (1.2) such as truncated polynomials or thin plate splines. We use B-splines basis due to their good properties. Basically, B-splines consist of polynomial pieces connected by a set of knots in a specific way. The general properties of a B-spline of order p, given in Eilers and Marx (1996), are the following:

- it consists of $p+1$ polynomial pieces, each of degree p;
- the polynomial pieces join at p inner knots;
- at joining points, derivatives up to order $p-1$ are continous;
- the B-spline is positive on a domain spanned by $p+2$ knots; everywhere else its zero;
- except at the boundaries, it overlaps with $2p$ polynomial pieces of its neighbors;
- at given x, $p+1$ B-splines are nonzero.

In practice, it is usual to divide the domain interval of $\boldsymbol{x}$ into k intervals with $k+1$ equally-spaced knots, covering each interval by $p+1$ B-splines of degree p, usually p is taken equal to 3. The number of B-splines in the regression basis (the number of columns of $\boldsymbol{B}$) is $c = k + p$.

Among the properties of the P-splines with bases B-splines, it is noteworthy that they do not suffer from edge effects, i.e., when the curve is extended outside the domain of $\boldsymbol{x}$, the curve does not fall quickly to 0 as it happens in the kernels. In addition, if the curve is a polynomial, the P-spline adjust it exactly. Moreover, they keep the moments, i.e., the mean and variance of the adjusted values will be the same as the data, regardless of the smoothing parameter. As it is reported in Eilers and Marx (2010) and Eilers et al. (2015) the number of knots is essentially irrelevant, as long as it is large enough.

In Figure 1.1 it can be seen B-splines of several degrees. In the top left figure it is shown a B-spline of degree 1, it consists of two linear pieces, in the top right figure it can be seen several B-splines of degree 1. In the bottom left figure a B-spline of degree 2 is

shown, and in the bottom right figure several B-splines of degree 2 are shown.

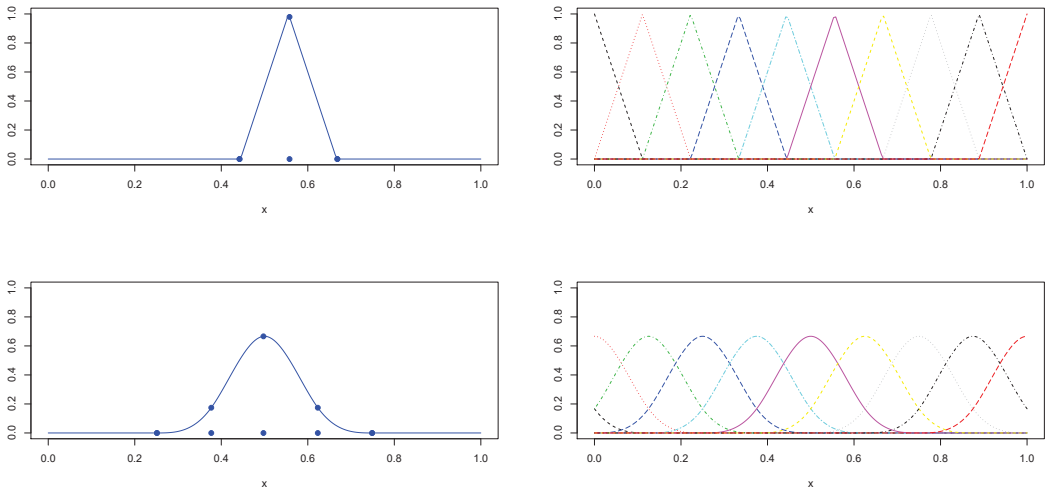


Figure 1.1: B-splines of degree 1 and 2.

Penalties

The penalty matrix $\boldsymbol{P}$ in (1.3) penalizes the difference between adjacent coefficients, it is a matrix of dimension $c \times c$ defined from the difference operator of order q, Δ^q:

$$\boldsymbol{P} = (\Delta^q)'\Delta^q. \tag{1.5}$$

The order of the penalty, q, controls the jumps between adjacent coefficients. For instance, if $q = 1$, $\boldsymbol{P}$ penalizes the difference between consecutive coefficients, and if $q = 2$, the penalty is equivalent to

$$(\theta_1 - 2\theta_2 + \theta_3)^2 + ... + (\theta_{c-2} - 2\theta_{c-1} + \theta_c)^2 = \boldsymbol{\theta}'\boldsymbol{D}_2'\boldsymbol{D}_2\boldsymbol{\theta},$$

where, for instance for $c = 4$, $\boldsymbol{D}_2 = \begin{bmatrix} 1 & -2 & 0 & 0 \\ 0 & 1 & -2 & 1 \end{bmatrix}$.
In general, for any order q, the penalty matrix, $\boldsymbol{P}$, has the form $\boldsymbol{D}_q'\boldsymbol{D}_q$, with $\boldsymbol{D}_q$ a difference matrix of order q and dimension $(c - q) \times c$.

To illustrate the P-splines approach, we have generated 200 points, (x_i, y_i), from the function $f(x_i) = 2 + \sin(2x_i) + 0.5\epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, 1)$ and $x_i \sim \text{Unif}[0, 1]$. Figure 1.2 illustrates the fit of the data with and without penalty. The left panel shows the fit

without penalty, i.e., $\lambda = 0$, the right panel shows the fit with $\lambda = 1$.

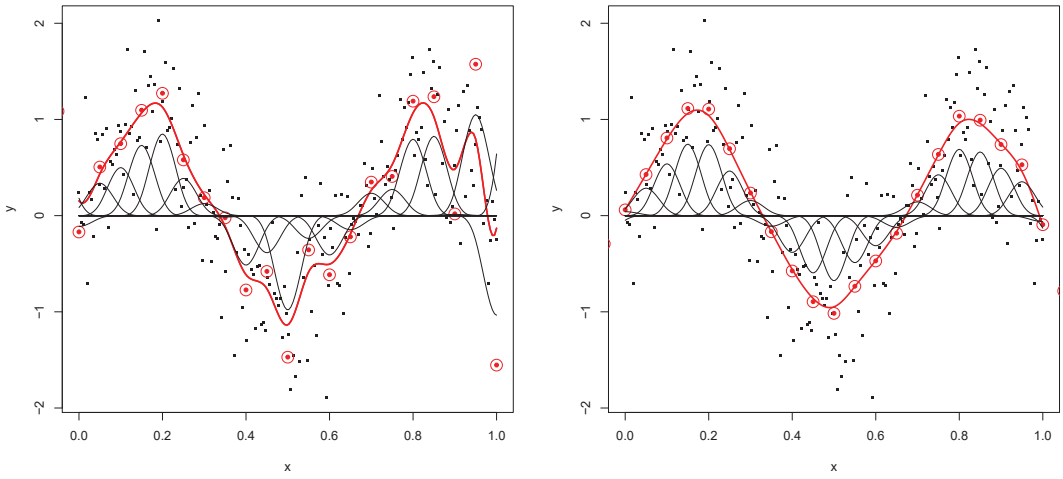


Figure 1.2: Left panel: B-splines with unpenalized coefficients. Right panel: B-splines with penalized coefficients.

Smoothing parameter selection

As it was said, the P-spline model fit requires the choice of the optimal smoothing parameter λ. There exist several methods to choose the optimal value of λ, which are based on: *cross-validation*, where the idea is to leave-out one observation in turn and then fit the model to the remaining data and calculate the square difference between the missing data and its prediction, or *information criterion*, which idea is to deal with the trade-off between the goodness of fit of the model and the complexity of the model, using also the effective dimension. Examples of these methods can be seen in Eilers and Marx (1996), some of them are:

- Ordinary cross-validation:

$$\mathrm{CV}(\lambda) = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2.$$

- Generalized cross-validation:

$$\mathrm{GCV}(\lambda) = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{n - \mathrm{trace}(\boldsymbol{H})} \right)^2.$$

- Information criterion methods:

$$\text{IC}(\lambda) = \text{Dev}(\boldsymbol{y}; \boldsymbol{\theta}, \lambda) + \delta \ \text{ED}(\boldsymbol{\theta}, \lambda),$$

  where the deviance is a measure of the quality of the fit, defined as $\text{Dev}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 2\{\mathcal{L}(\boldsymbol{y}) - \mathcal{L}(\hat{\boldsymbol{y}})\}$. For Gaussian data $\text{Dev}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. For non-Gaussian data, the deviance is based on a generalization of the sum of squares, and depends on the distributional assumptions. δ penalizes the effective dimension (ED), if $\delta = 2$ and $\delta = \log(n)$ the information criteria is known as Akaike information criteria and Bayesian information criteria, respectively.

In the previous equations $\boldsymbol{H}$ is the hat matrix, defined in (1.4), and h_{ii} its diagonal elements. The best λ is the value that minimizes $\text{CV}(\lambda)$, $\text{GCV}(\lambda)$ or $\text{IC}(\lambda)$.
In the next section, we will see the mixed model formulation of a P-spline model. Such formulation allows us to calculate λ as a variance component estimation problem, and then, we do not need to use some smoothing selection method to get an optimal smoothing parameter.

1.2.2 Mixed models

In this section, we illustrate the main ideas related to mixed models since rewriting a P-spline as a mixed model can be very useful due to, principally, two reasons: model building and computation. For instance, if we are working with an autoregressive model, rewriting the associated P-spline model as a mixed model, we can estimate, simultaneously, the smoothing parameter and the correlation parameter. We will start by giving a brief overview of the mixed models methodology.
Linear mixed effects models (LMMs) are an extension of regression models which incorporate random effects.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \text{ with } \boldsymbol{\alpha} \sim \mathcal{N}(0, \boldsymbol{G}) \text{ and } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{R}), \tag{1.6}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the model matrices and $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the fixed and random effects coefficients respectively. The random effects have covariance matrix $\boldsymbol{G}$, which depends on the variance of the random effects $\sigma^2_{\boldsymbol{\alpha}}$. Assuming that the errors are i.i.d., $\boldsymbol{R} = \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}$. For more details see Searle et al. (1992).

Below, it is shown how the fixed effects estimation and random effects prediction can be done and how a P-spline can be reformulated as a mixed model.

Parameter estimation

The Henderson's mixed model equations allow us to obtain the best linear unbiased estimator of $\boldsymbol{X\beta}$ and the best linear unbiased predictor of $\boldsymbol{Z\alpha}$. They are obtained maximizing the joint density of $\boldsymbol{y}$ and $\boldsymbol{\alpha}$:

$$f(\boldsymbol{y},\boldsymbol{\alpha}) = f(\boldsymbol{y}|\boldsymbol{\alpha})f(\boldsymbol{\alpha}), \quad \boldsymbol{y}|\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{X\beta}+\boldsymbol{Z\alpha},\boldsymbol{R}), \quad \boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{G}),$$

where $\boldsymbol{R}$ depends on $\sigma^2_{\boldsymbol{\epsilon}}$ and $\boldsymbol{G}$ on $\sigma^2_{\boldsymbol{\alpha}}$, the log-likelihood is:

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\alpha}) \propto -\frac{1}{2}\left[\log|\boldsymbol{R}| + \log|\boldsymbol{G}| + (\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Z\alpha})'\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Z\alpha}) + \boldsymbol{\alpha}'\boldsymbol{G}^{-1}\boldsymbol{\alpha}\right],$$

deriving respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ we obtain the equations of Henderson (1975):

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}. \tag{1.7}$$

The solutions are:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}, \tag{1.8}$$

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{G}\boldsymbol{Z}'\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{1.9}$$

where $\hat{\boldsymbol{V}} = \boldsymbol{Z}\hat{\boldsymbol{G}}\boldsymbol{Z}' + \hat{\boldsymbol{R}}$. Note that $\hat{\boldsymbol{V}}$ includes the variance components $\hat{\sigma}^2_{\boldsymbol{\epsilon}}$ and $\hat{\sigma}^2_{\boldsymbol{\alpha}}$, through the covariance matrices $\hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{G}}$, respectively.

The variance components can be estimated by maximum likelihood (ML) or by restricted or residual maximum likelihood (REML). Let us rewrite (1.6) as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}^*, \quad \text{where} \quad \boldsymbol{\epsilon}^* = \boldsymbol{Z\alpha} + \boldsymbol{\epsilon}.$$

The variance-covariance matrix of $\boldsymbol{y}$ is:

$$\boldsymbol{V} = \text{Cov}(\boldsymbol{y}) = \text{Cov}(\boldsymbol{\epsilon}^*) = \text{Cov}(\boldsymbol{Z\alpha}) + \text{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{Z}'\boldsymbol{G}\boldsymbol{Z} + \boldsymbol{R},$$

i.e., $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X\beta},\boldsymbol{V})$ and therefore:

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{V}) = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}).$$

Notice that the maximum log-likelihood estimator of $\boldsymbol{\beta}$ is the same as (1.8), substituting

that formula in $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{V})$, we obtain the following expression:

$$\mathcal{L}(\boldsymbol{V}) = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\boldsymbol{y}'(\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1})\boldsymbol{y}. \tag{1.10}$$

However, the maximum likelihood estimators of variance components are biased. Due to this fact, the most popular method to estimate the variance components is the restricted likelihood estimation (REML), that accounts for the degrees of freedom used for the fixed effects estimation. REML results from modifying the standard likelihood function using generalized least squares residuals, as suggested in Patterson and Thompson (1971), i.e., in REML estimation one maximizes de log-likelihood for the residual vector $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The restricted maximum log-likelihood is:

$$\mathcal{L}_{\text{REML}}(\boldsymbol{V}) = \mathcal{L}(\boldsymbol{V}) - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|.$$

In both cases, $\boldsymbol{V}$ is obtained computing the maximum of $\mathcal{L}(\boldsymbol{V})$ and $\mathcal{L}_{\text{REML}}(\boldsymbol{V})$, respectively, using numerical methods. For practical purposes, in the case of i.i.d. errors for fast estimation of the covariance components, we use the algorithm of Schall (1991), that is generalized and implemented in the context of smoothing in Rodríguez-Álvarez et al. (2018).

Mixed models formulation for P-splines

The connection between penalized smoothing and mixed models was established thirty years ago in Green (1987) (see also Currie and Durbán 2002 and Wand 2003). The key point of this equivalence is the fact that the smoothing parameter becomes a ratio of variances, $\lambda = \sigma^2_{\boldsymbol{\epsilon}}/\sigma^2_{\boldsymbol{\alpha}}$, and both variance components can be estimated through restricted maximum likelihood procedure (REML) (see Patterson and Thompson 1971). Therefore, it is not longer necessary to estimate λ via a cross-validation method or an information criterion. The idea is extensively discussed in Ruppert et al. (2003).

This representation allows us to include smoothing in a large class of models and the use of the methodology and software already developed for mixed models for estimation and inference. In the B-spline basis context, Currie and Durbán (2002) extended the P-spline model formulation into mixed model.

To represent a penalized spline model as a mixed model it is necessary to find a new basis that allows the representation of model (1.2) with its associated penalty as a mixed model (1.6). This representation decomposes the fitted values as the sum of a unpenalized polynomial part, $\boldsymbol{X}\boldsymbol{\beta}$, and a penalized non-linear smooth term $\boldsymbol{Z}\boldsymbol{\alpha}$.

To rewrite a P-spline model as a mixed model, we have to set a transformation matrix $\boldsymbol{\Omega}$ such that:

$$\boldsymbol{B\Omega} = [\boldsymbol{X} \mid \boldsymbol{Z}] \text{ and } \boldsymbol{\Omega'\theta} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} \text{ to have } \boldsymbol{B\theta} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha},$$

where $\boldsymbol{\Omega}$ is an orthogonal matrix. Notice that the transformation matrix, $\boldsymbol{\Omega}$, can be splitted into two matrices: one associated to the fixed part, $\boldsymbol{\Omega}_f$, and another associated to the random part, $\boldsymbol{\Omega}_r$, i.e. $\boldsymbol{\Omega} = [\boldsymbol{\Omega}_f \mid \boldsymbol{\Omega}_r]$. The fixed effects are not penalized, therefore $\boldsymbol{X}$ may be any matrix verifying $[\boldsymbol{X} \mid \boldsymbol{Z}]$ has full rank. For the choice of $\boldsymbol{\Omega}_r$ there are different options, following the proposal of Currie and Durbán (2002) and Lee (2010), we use the singular value decomposition (SVD) of the penalty matrix:

$$\boldsymbol{D}_q'\boldsymbol{D}_q = [\boldsymbol{U}_f \mid \boldsymbol{U}_r] \begin{bmatrix} \boldsymbol{O}_q & \\ & \tilde{\boldsymbol{\Sigma}} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_f' \\ \boldsymbol{U}_r' \end{bmatrix},$$

where $\boldsymbol{U}_f$ of dimension $c \times q$ contains the eigenvectors associated to the q zero eigenvalues, $\boldsymbol{U}_r$ of dimension $c \times (c-q)$ contains the eigenvectors associated to the non-zero eigenvalues, $\boldsymbol{O}_q$ is a null matrix of dimension $q \times q$ and $\tilde{\boldsymbol{\Sigma}}$ is a diagonal matrix containing the non-zero eigenvalues of dimension $(c-q) \times (c-q)$.

For penalty of order q, one can take the design matrix of a polynomial of order $q-1$ as the fixed effect matrix, i.e.,

$$\boldsymbol{X} = \left[\mathbf{1}_n \mid \boldsymbol{x}_i \mid \boldsymbol{x}_i^2 \mid ... \mid \boldsymbol{x}_i^{q-1}\right],$$

where $\mathbf{1}_n$ is a column vector of ones. And the random effects matrix, the following matrix defined from the SVD of the penalty matrix:

$$\boldsymbol{Z} = \boldsymbol{B\Omega}_r \text{ with } \boldsymbol{\Omega}_r = \boldsymbol{U}_r\tilde{\boldsymbol{\Sigma}}^{-1/2}, \text{ of dimension } c \times (c-q).$$

Notice that for the given transformation, the penalty $\lambda\boldsymbol{\theta}'\boldsymbol{P\theta}$ in the mixed model framework has the following form:

$$\lambda\boldsymbol{\theta}'\boldsymbol{P\theta} = \boldsymbol{\theta}'\lambda\boldsymbol{D}_q'\boldsymbol{D}_q\boldsymbol{\theta} = \lambda\boldsymbol{\theta}' \left[\boldsymbol{U}_f \mid \boldsymbol{U}_r\right] \begin{bmatrix} \boldsymbol{O}_q & \\ & \tilde{\boldsymbol{\Sigma}} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_f' \\ \boldsymbol{U}_r \end{bmatrix} \boldsymbol{\theta} = \lambda\boldsymbol{\alpha}'\boldsymbol{I}_{c-q}\boldsymbol{\alpha}.$$

Therefore, for the design model matrices, $\boldsymbol{X}$ and $\boldsymbol{Z}$, and the penalty, $\lambda\boldsymbol{\alpha}'\boldsymbol{I}_{c-q}\boldsymbol{\alpha}$, the

penalized sum of squares (1.3) becomes:

$$S(\boldsymbol{\beta}, \boldsymbol{\alpha}; \lambda) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\alpha})' \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\alpha}) + \frac{1}{\sigma^2_{\boldsymbol{\alpha}}} \boldsymbol{\alpha}' \boldsymbol{I}_{c-q} \boldsymbol{\alpha}, \tag{1.11}$$

and, if we differentiate (1.11) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, it is straightforward to obtain the standard mixed model equations in (1.8) and (1.9). Notice that with this reparametrization, the variance components matrix is $\boldsymbol{G} = \sigma^2_{\boldsymbol{\alpha}} \boldsymbol{I}_{c-q}$.

Once we have given a brief summary of the P-splines methodology and introduced the basic notation, we finish this chapter with an overview of the thesis.

1.3 Dissertation structure

This thesis is organized as follows, in Chapter 2 we give a general framework to predict out-of-sample values with smooth models that do not include interaction terms. To that end, we unify proposals available in the literature, in contexts such as penalized regression, mixed models or global optimization, for prediction with any regression basis and quadratic penalty. For the particular case of penalties based on differences between adjacent coefficients, we prove the equivalence of all methods and show some properties of the coefficients that determine the prediction. We also introduce the concept of "memory of a P-spline" as a tool to know how much of the known information we are using to predict. The illustration of the methodology is shown with several real datasets. In Chapter 3, we study the methodology proposed in Currie et al. (2004) for the case of P-splines models with interaction terms and extend it to the mixed model framework and to obtain predictions when more than one independent covariate is extended. In the multidimensional setting, we prove that the fit is not invariant to out-of-sample prediction and propose the use of restrictions to overcome this problem. In order to illustrate how restrictions can be used in different situations, we show how to solve the crossover problem of adjacent ages when mortality tables are forecasted. In Chapter 4 we widen the application of the methodology proposed for the case of Smooth-ANOVA models (models that allow us to include interaction terms that can be decomposed as a sum of several smooth functions). To illustrate the methodology, we analyze data coming from a forestry trial with the aim of predicting. We also compare the performance of 2D interaction models and Smooth-ANOVA models (both with and without imposing invariance of the fit) though a simulation study. In Chapter 5, we extend the developed methodology for generalized linear models (GLMs), where available approaches for estimation in the GLMs framework are adapted to fit and predict simultaneously. Finally, in Chapter 6 we summarize the main contributions given in this thesis and suggest possible future lines

for research.

Chapter 2

Prediction of new observations in additive P-spline models

This chapter develops the approach for prediction with penalized splines to the case in which the response variable is Gaussian and there are only additive smooth terms. The structure of the chapter is the following, in Section 2.1 we introduce the approach prosed by Currie et al. (2004) for univariate Gaussian data and, for the particular case of penalties based on differences (Eilers and Marx 1996), we demonstrate several properties of the new coefficients in terms of the order of the penalty. Section 2.2 is dedicated to obtain predictions in the context of mixed models through different methodologies: i) we extend the approach proposed in Gilmour et al. (2004) to predict in the case of smooth mixed models, and we connect it to the theory of conditional distributions, which allows us to compute prediction intervals, ii) we extend the approach proposed by Currie et al. (2004) to consider the prediction of new observations in the mixed model framework, and iii) we develop a method to predict in penalized regression based on the method proposed in Sacks et al. (1989). In the context of penalties based on differences between adjacent coefficients, the equivalence of the different methods is shown. The methodology is illustrated in Section 2.3 with three real data sets: the first one allows us to show an example of predicting within the framework of additive models and in the case when out-of-sample prediction is needed to the left and right of the interval where the covariate is observed. The second dataset illustrates a classic example where forecasting is needed, when data are collected over time, and the third dataset shows an example where the errors are correlated. Although, as we have shown all the methods give us the same result, in order to obtain the prediction intervals we use the two-stage approach.

2.1 Prediction with smooth models and quadratic penalties

Based on the idea of the P-slines and with the aim of proposing a general methodology to predict in penalized regression, in this section we present the method that allow us

to estimate, nonparametrically, the smooth curve and to predict new observations. If the prediction is within sample everything is straightforward, the coefficients remain the same and we just have to extend the basis to include the points in which we want to predict. In the case of out-of-sample prediction we obviously have to extend the basis but also the penalty to penalize the new coefficients.

Currie et al. (2004) proposed a method to fit and predict simultaneously in penalized regression models. We call their proposal "*the missing value approach*" subsequently and give a brief summary of their methodology.

In the framework of model (1.1), given a vector $\boldsymbol{y}$ of n observations of the response variable, suppose that we want to predict n_p new values $\boldsymbol{y}_p$ at $\boldsymbol{x}_p$, where $\boldsymbol{x}_p$ may be within or, more interestingly, outside of the range of observed values x_i. In the following we focus on the case when $\boldsymbol{x}_p$ is not in the convex hull of x_i, $i = 1, ..., n$. We define the new vector of observations as

$$\boldsymbol{y}_+ = (\boldsymbol{y}', \boldsymbol{y}_p')', \tag{2.1}$$

which contains the observed response $\boldsymbol{y}$ and the unknown values $\boldsymbol{y}_p$ to be predicted. A new extended B-spline basis, $\boldsymbol{B}_+$, is built from a new set of knots that consists of the original knots covering x_i, $i = 1, ..., n$, and extended to the range of the n_p values of x_{p_j}, $j = 1, ..., n_p$. This leads to the basis:

$$\boldsymbol{B}_+ = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{O} \\ \boldsymbol{B}_1 & \boldsymbol{B}_2 \end{bmatrix} \text{ of size } n_+ \times c_+, \tag{2.2}$$

where $\boldsymbol{B}$ is the $n \times c$ basis used for fitting the trend component, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are auxiliary bases for prediction up to $n_+ = n + n_p$ values, which are of sizes $n_p \times c$ and $n_p \times c_p$, respectively, and $c_+ = c + c_p$. Figure 2.1 represents an extended splines basis. We show the original basis $\boldsymbol{B}$ in black, the $\boldsymbol{B}_1$ component in grey and the $\boldsymbol{B}_2$ part with dashed line. In Appendix A.1 we show the code that allows to build $\boldsymbol{B}_+$.

Associated to the new basis $\boldsymbol{B}_+$, a new vector of coefficients, $\boldsymbol{\theta}_+ = (\boldsymbol{\theta}', \boldsymbol{\theta}_p')'$, is defined, with length $c_+ \times 1$. A new quadratic penalty associated to the new set of coefficients needs to be introduced, let say $\boldsymbol{P}_+$. Similar to $\boldsymbol{B}_+$, we can also decompose $\boldsymbol{P}_+$ to

$$\boldsymbol{P}_+ = \begin{bmatrix} \boldsymbol{P}_1 & \boldsymbol{P}_2 \\ \boldsymbol{P}_2' & \boldsymbol{P}_3 \end{bmatrix}. \tag{2.3}$$

In the case that $\boldsymbol{P}_+$ is built from q-th order difference matrices, i.e. $\boldsymbol{P}_+ = \boldsymbol{D}'_+\boldsymbol{D}_+$, we have

$$\boldsymbol{D}_+ = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{O} \\ \boldsymbol{D}_1 & \boldsymbol{D}_2 \end{bmatrix} \tag{2.4}$$

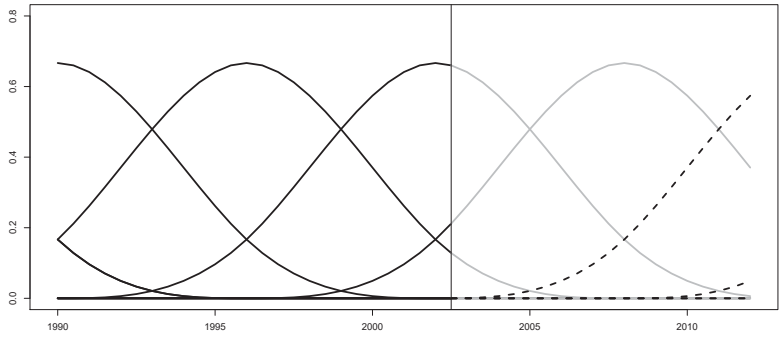


Figure 2.1: Example of an extended basis to the right of the data (forward).

of order q and size $(c_+ - q) \times c_+$, with $\boldsymbol{D}$ the difference matrix used to build the penalty matrix for the observed data. Moreover, for instance for a second order penalty, $\boldsymbol{D}_+$ is a banded matrix with three non-zero elements per row. Notice that in this case the extended penalty matrix is

$$\boldsymbol{P}_+ = \begin{bmatrix} \boldsymbol{D}'\boldsymbol{D} + \boldsymbol{D}_1'\boldsymbol{D}_1 & \boldsymbol{D}_1'\boldsymbol{D}_2 \\ \boldsymbol{D}_2'\boldsymbol{D}_1 & \boldsymbol{D}_2'\boldsymbol{D}_2 \end{bmatrix}, \tag{2.5}$$

i.e., $\boldsymbol{P}_1 = \boldsymbol{D}'\boldsymbol{D} + \boldsymbol{D}_1'\boldsymbol{D}_1$, $\boldsymbol{P}_2 = \boldsymbol{D}_1'\boldsymbol{D}_2$ and $\boldsymbol{P}_3 = \boldsymbol{D}_2'\boldsymbol{D}_2$. Here the subscripts do not indicate the order of the penalty but the blocks of the extended differences matrix.

The model can be fitted and predicted simultaneously by minimizing the following penalized least squares criterion for $\boldsymbol{\theta}_+$:

$$S \;=\; (\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_+)'\tilde{\boldsymbol{R}}_+^{-1}(\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_+) + \lambda\boldsymbol{\theta}_+'\boldsymbol{P}_+\boldsymbol{\theta}_+, \tag{2.6}$$

where the unknown $\boldsymbol{y}_p$ values of $\boldsymbol{y}_+$ are arbitrary and $\tilde{\boldsymbol{R}}_+^{-1}$, with dimension $n_+ \times n_+$, is the inverse of

$$\frac{1}{\sigma_\epsilon^2}\begin{bmatrix} \boldsymbol{R} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{R}_{pp} \end{bmatrix},$$

with $\boldsymbol{R}$ the variance covariance-matrix of the errors associated to the observed data and $\boldsymbol{R}_{pp}$ a diagonal matrix with entries infinite to expresses that we do not have any information about the data $\boldsymbol{y}_p$. Notice that $\tilde{\boldsymbol{R}}_+^{-1} = \begin{bmatrix} \tilde{\boldsymbol{R}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}$, and that for the particular case of i.i.d. errors, $\tilde{\boldsymbol{R}}_+^{-1}$ is a diagonal matrix of dimension $n_+ \times n_+$ with 0 entries if the data is missing, that is for $\boldsymbol{y}_p$, and 1 if the data is observed, that is for $\boldsymbol{y}$. Differentiating

with respect to $\boldsymbol{\theta}_+$ leads to

$$\frac{\partial S}{\partial \boldsymbol{\theta}_+} = -2\boldsymbol{B}_+'\tilde{\boldsymbol{R}}_+^{-1}(\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_+) + 2\lambda \boldsymbol{P}_+\boldsymbol{\theta}_+. \tag{2.7}$$

The penalized least square solution is given by:

$$\hat{\boldsymbol{\theta}}_+ = (\boldsymbol{B}_+'\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{B}_+ + \lambda \boldsymbol{P}_+)^{-1}\boldsymbol{B}_+'\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{y}_+,$$

and $\hat{\boldsymbol{y}}_+ = \boldsymbol{H}_+\boldsymbol{y}_+$ with $\boldsymbol{H}_+ = \boldsymbol{B}_+(\boldsymbol{B}_+'\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{B}_+ + \lambda \boldsymbol{P}_+)^{-1}\boldsymbol{B}_+'\tilde{\boldsymbol{R}}_+^{-1}$. Note that $\lambda > 0$ is required so that the matrix inversion in (2.1) exists.

Notice that the same expressions for the coefficients are obtained directly by optimizing with respect to $\boldsymbol{\theta}_+$ and $\boldsymbol{y}_p$ the augmented least squares problem,

$$S_+ = (\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_+)'\tilde{\boldsymbol{R}}_+^{-1}(\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_+) + \lambda \boldsymbol{\theta}_+\boldsymbol{P}_+\boldsymbol{\theta}_+, \tag{2.8}$$

since, defining $\boldsymbol{B}_o = [\boldsymbol{B} \mid \boldsymbol{O}]$ and $\boldsymbol{B}_p = [\boldsymbol{B}_1 \mid \boldsymbol{B}_2]$, (2.8) can be written as

$$\begin{bmatrix} \boldsymbol{y} - \boldsymbol{B}_o\boldsymbol{\theta}_+ \\ \boldsymbol{y}_p - \boldsymbol{B}_p\boldsymbol{\theta}_+ \end{bmatrix}' \begin{bmatrix} \tilde{\boldsymbol{R}}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{y} - \boldsymbol{B}_o\boldsymbol{\theta}_+ \\ \boldsymbol{y}_p - \boldsymbol{B}_p\boldsymbol{\theta}_+ \end{bmatrix} + \lambda \boldsymbol{\theta}_+'\boldsymbol{P}_+\boldsymbol{\theta}_+,$$

and taking derivatives with respect to $\boldsymbol{y}_p$, we get

$$\hat{\boldsymbol{y}}_p = \boldsymbol{B}_p\hat{\boldsymbol{\theta}}_+. \tag{2.9}$$

Moreover, if $\hat{\boldsymbol{\theta}}_+$ is the vector that minimizes (2.8), it verifies:

$$\boldsymbol{B}_+'\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{B}_+\hat{\boldsymbol{\theta}}_+ + \lambda \boldsymbol{P}_+\hat{\boldsymbol{\theta}}_+ = \boldsymbol{B}_+'\boldsymbol{y}_+,$$

that is,

$$\boldsymbol{B}_o'\tilde{\boldsymbol{R}}^{-1}\boldsymbol{B}_o\hat{\boldsymbol{\theta}}_+ + \boldsymbol{B}_p'\boldsymbol{B}_p\hat{\boldsymbol{\theta}}_+ + \lambda \boldsymbol{P}_+\hat{\boldsymbol{\theta}}_+ = \boldsymbol{B}_o'\tilde{\boldsymbol{R}}^{-1}\boldsymbol{y} + \boldsymbol{B}_p'\boldsymbol{y}_p. \tag{2.10}$$

Now, using (2.9) to rewrite (2.10) we get, after some simplification:

$$\boldsymbol{B}_o'\tilde{\boldsymbol{R}}^{-1}\boldsymbol{B}_o\hat{\boldsymbol{\theta}}_+ + \lambda \boldsymbol{P}_+\hat{\boldsymbol{\theta}}_+ = \boldsymbol{B}_o'\tilde{\boldsymbol{R}}^{-1}\boldsymbol{y}.$$

Hence $\hat{\boldsymbol{\theta}}_+$ equals

$$\hat{\boldsymbol{\theta}}_+ = (\boldsymbol{B}_o'\tilde{\boldsymbol{R}}^{-1}\boldsymbol{B}_o + \lambda \boldsymbol{P}_+)^{-1}\boldsymbol{B}_o'\tilde{\boldsymbol{R}}^{-1}\boldsymbol{y} = (\boldsymbol{B}_+'\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{B}_+ + \lambda \boldsymbol{P}_+)^{-1}\boldsymbol{B}_+'\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{y}_+.$$

Writing the fit and the prediction as a function of the extended penalty matrix (2.3) and

applying Theorem 9.6.1 given in Harville (2000), we get

$$\hat{\boldsymbol{y}}_+ = \boldsymbol{B}_+ \begin{bmatrix} \boldsymbol{I} \\ -\boldsymbol{P}_3^- \boldsymbol{P}_2' \end{bmatrix} (\boldsymbol{B}' \tilde{\boldsymbol{R}}^{-1} \boldsymbol{B} + \lambda \boldsymbol{P}_1 - \lambda \boldsymbol{P}_2 \boldsymbol{P}_3^- \boldsymbol{P}_2')^{-1} \boldsymbol{B}' \tilde{\boldsymbol{R}}^{-1} \boldsymbol{y}, \tag{2.11}$$

where the superscript $(^-)$ denotes the generalized inverse for the cases in which the inverse does not exist. Note that $\hat{\boldsymbol{y}}_+ = (\hat{\boldsymbol{y}}', \hat{\boldsymbol{y}}_p')'$ is a fitted mean value, i.e. $\hat{\boldsymbol{y}}_+ = \widehat{\mathbb{E}[\boldsymbol{y}_+]}$. Hence, in particular the new values $\boldsymbol{y}_p$ are predicted by their fitted mean, where the fit is based on the observed values $\boldsymbol{y}$. Taking formula (2.11) we can derive the expectation of $\hat{\boldsymbol{y}}_p$, the new values, yielding

$$\mathbb{E}[\hat{\boldsymbol{y}}_p] = (\boldsymbol{B}_1 - \boldsymbol{B}_2 \boldsymbol{P}_3^- \boldsymbol{P}_2')(\boldsymbol{B}' \tilde{\boldsymbol{R}}^{-1} \boldsymbol{B} + \lambda \boldsymbol{P}_1 - \lambda \boldsymbol{P}_2 \boldsymbol{P}_3^- \boldsymbol{P}_2)^{-1} \boldsymbol{B}' \tilde{\boldsymbol{R}}^{-1} \boldsymbol{B} \boldsymbol{\theta}.$$

2.1.1 Properties of the predictions in the case of P-splines with penalties based on differences

When penalties are based on differences between adjacent coefficients, for interpolating, there is no need for new coefficients and the B-spline coefficients form a polynomial sequence of degree $2q-1$ (Eilers and Marx 2010), for instance when $q = 2$, we get cubic interpolation. If we extrapolate the method above satisfies certain important properties. The subsequent results are based on a basis constructed from equally spaced knots, however, the results extend also to the non-equal spaced knots case, if we define the appropriately scaled penalty matrices. The central results are the following:

i) The fit remains the same regardless of the prediction horizon (i.e. the coefficients that yield the fit do not change).

ii) The shape of the prediction is determined by the order of the penalty.

These properties are an immediate consequence of the following theorems.

Theorem 2.1. *The coefficients from minimizing (2.6) with extended penalty matrix (2.5) satisfy the following properties:*

I. The first c coefficients of $\hat{\boldsymbol{\theta}}_+$, are those obtained from the fit of $\boldsymbol{y}$, i.e.:

$$\hat{\boldsymbol{\theta}}_{+_{1,\ldots,c}} = \hat{\boldsymbol{\theta}}.$$

II. The coefficients for the n_p predicted values are $\hat{\boldsymbol{\theta}}_p = -\boldsymbol{D}_2^{-1} \boldsymbol{D}_1 \hat{\boldsymbol{\theta}}$.

Proof. Substituting the blocks of $\boldsymbol{P}_+$ by their specific values in (2.11) we have that:

$$\hat{\boldsymbol{y}}_+ = \boldsymbol{B}_+ \begin{bmatrix} \boldsymbol{I} \\ -\boldsymbol{D}_2'\boldsymbol{D}_1 \end{bmatrix} (\boldsymbol{B}'\tilde{\boldsymbol{R}}^{-1}\boldsymbol{B} + \lambda \boldsymbol{P})^{-1}\boldsymbol{B}'\tilde{\boldsymbol{R}}^{-1}\boldsymbol{y}$$

i.e., the first c coefficients of $\hat{\boldsymbol{\theta}}_+$ are:

$$(\boldsymbol{B}'\tilde{\boldsymbol{R}}^{-1}\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\boldsymbol{y},$$

the same as the ones that give the fit, and the additional coefficients are:

$$\hat{\boldsymbol{\theta}}_p = -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1(\boldsymbol{B}'\tilde{\boldsymbol{R}}^{-1}\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\tilde{\boldsymbol{R}}^{-1}\boldsymbol{y} = -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1\hat{\boldsymbol{\theta}}. \tag{2.12}$$

■

If the knots are not equally-spaced the expression above would be modified since the penalty would have to account for the difference between the knots (Eilers and Marx 2010).

Corollary 2.1 (Theorem 2.1)**.** *Given penalties of order q, the new coefficients are combinations of order $q-1$ of the last q fitted coefficients.*

Proof. As the most popular penalties are of second or third order, the proof of the previous corollary for such cases and for penalties of order 1 is shown.

• Differences of order 1.
Suppose a difference matrix with first order penalty $\boldsymbol{D}_+$ of dimensions $(c_+ - 1) \times c_+$,

$$\boldsymbol{D}_+ = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{O} \\ \boldsymbol{D}_1 & \boldsymbol{D}_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix},$$

where $\boldsymbol{D}_1$ has dimension $c_p \times c$, with c_p the additional number of parameters in $\boldsymbol{\theta}_+$, and

$\boldsymbol{D}_2$ has dimension $c_p \times c_p$:

$$\boldsymbol{D}_1 = \begin{bmatrix} 0 & 0 & \cdots & 0 & -1 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \quad \boldsymbol{D}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & -1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}.$$

Then, the additional vector of coefficients in (2.12) is:

$$\hat{\boldsymbol{\theta}}_p = -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1\hat{\boldsymbol{\theta}} = -\begin{bmatrix} 0 & 0 & \cdots & 0 & -1 \\ \vdots & \vdots & \cdots & 0 & -1 \\ \vdots & \vdots & \cdots & 0 & -1 \\ 0 & 0 & \cdots & 0 & -1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_{c-1} \\ \hat{\theta}_c \end{bmatrix} = \hat{\theta}_c \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix}.$$

Therefore, using differences of order 1 the new coefficients are equal to the last coefficient.

• Differences of order 2.
Suppose a difference matrix with second order penalty $\boldsymbol{D}_+$ of dimensions $(c_+ - 2) \times c_+$,

$$\boldsymbol{D}_+ = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{O} \\ \boldsymbol{D}_1 & \boldsymbol{D}_2 \end{bmatrix} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & -2 & 1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix},$$

where $\boldsymbol{D}_1$ has dimension $c_p \times c$, with c_p the additional number of parameters in $\boldsymbol{\theta}_+$, and $\boldsymbol{D}_2$ has dimension $c_p \times c_p$:

$$\boldsymbol{D}_1 = \begin{bmatrix} 0 & 0 & \cdots & 1 & -2 \\ 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \quad \boldsymbol{D}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ -2 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}.$$

Then, the additional vector of coefficients in (2.12) is:

$$\hat{\boldsymbol{\theta}}_p = -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1\hat{\boldsymbol{\theta}} = -\begin{bmatrix} 0 & 0 & \cdots & 1 & -2 \\ \vdots & \vdots & \cdots & 2 & -3 \\ \vdots & \vdots & \cdots & 3 & -4 \\ 0 & 0 & \cdots & 4 & -5 \\ \vdots & \vdots & \cdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_{c-1} \\ \hat{\theta}_c \end{bmatrix} = \hat{\theta}_c \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} + (\hat{\theta}_c - \hat{\theta}_{c-1}) \begin{bmatrix} 1 \\ 2 \\ 3 \\ \vdots \end{bmatrix}.$$

Therefore, using differences of order 2 the new coefficients are a linear combination of the last two coefficients obtained after fitting the observed data.

• Differences of order 3.
Suppose a difference matrix with third order penalty, $\boldsymbol{D}_+$ of dimensions $(c_+ - 3) \times c_+$,

$$\boldsymbol{D}_+ = \begin{bmatrix} -1 & 3 & -3 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & -1 & 3 & -3 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 3 & -3 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & -1 & 3 & -3 & 1 \end{bmatrix}.$$

In this case, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are:

$$\boldsymbol{D}_1 = \begin{bmatrix} 0 & 0 & 0 & \cdots & -1 & 3 & -3 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 3 \\ 0 & 0 & 0 & \cdots & 0 & 0 & -1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{D}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -3 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 3 & -3 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 3 & -3 & 1 \end{bmatrix}.$$

Therefore, by (2.12):

$$\hat{\boldsymbol{\theta}}_p \quad = \quad -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1\hat{\boldsymbol{\theta}} = -\begin{bmatrix} 0 & \cdots & 0 & -1 & 3 & -3 \\ 0 & \cdots & 0 & -3 & 8 & -6 \\ 0 & \cdots & 0 & -6 & 15 & -10 \\ 0 & \cdots & 0 & -10 & 24 & -15 \\ 0 & \cdots & 0 & -15 & 35 & -21 \\ 0 & \cdots & 0 & -21 & 48 & -28 \\ 0 & \cdots & 0 & -28 & 63 & -36 \\ 0 & \cdots & 0 & -36 & 80 & -45 \\ 0 & \cdots & 0 & -45 & 99 & -55 \\ \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\theta}_3 \\ \vdots \\ \hat{\theta}_{c-2} \\ \hat{\theta}_{c-1} \\ \hat{\theta}_c \end{bmatrix}$$

$$= \hat{\theta}_c \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} + \frac{3\hat{\theta}_c - 4\hat{\theta}_{c-1} + \hat{\theta}_{c-2}}{2} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ \vdots \end{bmatrix} + \frac{\hat{\theta}_c - 2\hat{\theta}_{c-1} + \hat{\theta}_{c-2}}{2} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ \vdots \end{bmatrix}^2 ,$$

in this case, the new coefficients are a linear combination of the last three coefficients obtained after fitting the observed values. The prediction is a quadratic polynomial.

■

In many situations, the fit is not greatly affected by the order of the penalty. However, there is an immediate connection between the penalty (or prior distribution) on the coefficients and the shape of the out-of-sample prediction shown in the above corollary. This is known in the framework of Bayesian P-splines, where the difference penalty corresponds to assuming a random walk prior on the coefficients (see Lang and Brezger 2004), however it is not common knowledge in non-Bayesian context. We believe this is an important result that needs to be addressed when using this methodology.

Although mortality data are often analyzed through a Poisson distribution, for the simple purpose of illustrating the result of Corollary 2.1 we use a data set on the log mortality rates of Spanish men aged 73 considering the log mortality rates as normal data. We use data from the Human Mortality Database (2018), over the period 1960-2016. In order to predict the log mortality rates of Spanish men aged 73 between 2016 and 2026, we apply the proposed methodology with B-splines of degree 3 as basis and three different penalty orders (1, 2 and 3). As it can be seen in Figure 2.2, if the penalty has order 1, the forecast is constant, if the penalty is of order 2 the forecast is linear and if penalty is of third order the forecast is quadratic.

2.2 Prediction with mixed-effects smooth models

We have established the connection between mixed models and P-splines in Section 1.2.2, so we can use the existing methodology in the context of mixed models to obtain predictions in penalized regression. Prediction in the context of mixed models is always done as a two-stage procedure: First fit and then predict. We show how to use the existing results in the context of smooth mixed models, and then, we will propose an alternative one-stage approach. For simplicity, we consider i.i.d. errors, i.e. $\boldsymbol{R} = \sigma_\epsilon^2 \boldsymbol{I}$.

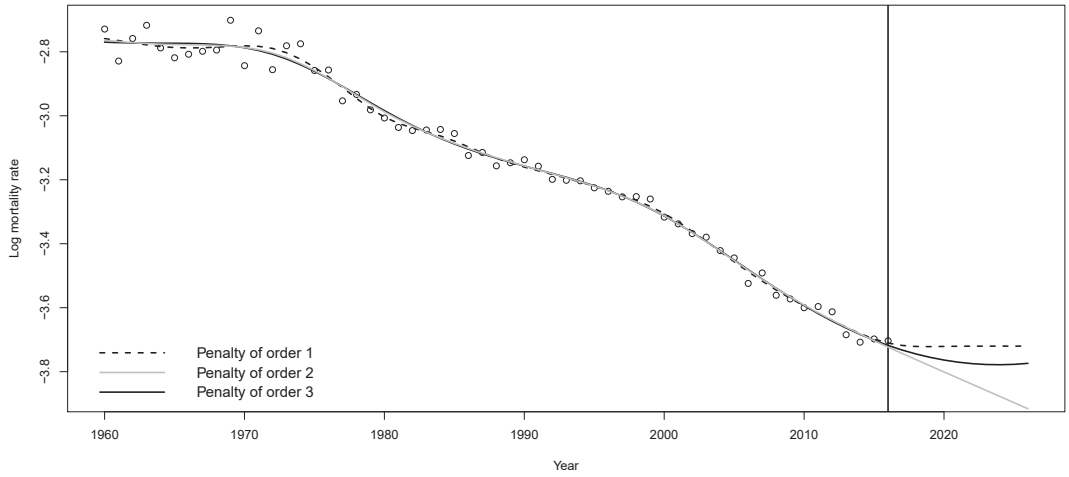


Figure 2.2: Fit and forecast result of applying the proposed methodology with penalty orders 1, 2 and 3 of a data set on the log mortality rates of Spanish men aged 73 between 1960 and 2016.

2.2.1 Two-stage approaches

Standard methodology for prediction

Gilmour et al. (2004) propose a method to predict new observations in which the prediction is a linear function of the best linear unbiased predictor (BLUP) of random effects and the best linear unbiased estimator (BLUE) of the fixed effects in the model. The results are based on the following augmented mixed model,

$$\boldsymbol{y}_+ = \boldsymbol{X}_+\boldsymbol{\beta} + \boldsymbol{Z}_+\boldsymbol{\alpha}_+ + \boldsymbol{\epsilon}_+, \tag{2.13}$$

i.e.:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{X}_p \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{Z} & \boldsymbol{O} \\ \boldsymbol{Z}_1 & \boldsymbol{Z}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}_p \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon}_p \end{bmatrix},$$

with $(\boldsymbol{\epsilon}', \boldsymbol{\epsilon}_p')' \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}_+)$, $\boldsymbol{\beta}$ the fixed effects (the same as the ones that give the fit, the fixed part is linear and therefore no new parameters for the linear part are needed), and $\boldsymbol{\alpha}_+ = (\boldsymbol{\alpha}', \boldsymbol{\alpha}_p')'$ the augmented random effects with covariance matrix

$$\text{Var}[\boldsymbol{\alpha}_+] = \boldsymbol{G}_+ = \begin{bmatrix} \boldsymbol{G} & \boldsymbol{G}_{op} \\ \boldsymbol{G}_{po} & \boldsymbol{G}_{pp} \end{bmatrix},$$

where $\boldsymbol{G}$ is the covariance matrix of the random effects in the model for the observed data, $\boldsymbol{G}_{op}$ is the covariance matrix between the random effects for the observed data

and for the unobserved data and $\boldsymbol{G}_{pp}$ is the covariance matrix of random effects for the unobserved data. The variance components, $\sigma^2_{\boldsymbol{\epsilon}}$ and $\sigma^2_{\boldsymbol{\alpha}}$, are estimated in the fit through restricted maximum likelihood procedure (REML) (Patterson and Thompson 1971).
Now we need to formulate the extended P-spline model into the extended mixed model (2.13). For that, we need to define an extended transformation matrix

$$\boldsymbol{\Omega}_{r_+} = \begin{bmatrix} \boldsymbol{\Omega}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Omega}_{p_r} \end{bmatrix},$$

where $\boldsymbol{\Omega}_r$ is the transformation matrix used for the observed data, and $\boldsymbol{\Omega}_{p_r}$ the one for the predicted values. Hence we have $\boldsymbol{Z}_1 = \boldsymbol{B}_1\boldsymbol{\Omega}_r$ and $\boldsymbol{Z}_2 = \boldsymbol{B}_2\boldsymbol{\Omega}_{p_r}$. There are many ways in which $\boldsymbol{\Omega}_{r_+}$ may be chosen. In the context of penalties based on differences, for simplicity, we chose $\boldsymbol{\Omega}_r = \boldsymbol{U}_r\boldsymbol{\Sigma}^{-1/2}$, based on the SVD of $\boldsymbol{D}'\boldsymbol{D} = \boldsymbol{U}\tilde{\boldsymbol{\Sigma}}\boldsymbol{U}'$, and $\boldsymbol{\Omega}_{p_r} = \boldsymbol{D}_2^{-1}$, with $\boldsymbol{D}$ and $\boldsymbol{D}_2$ blocks of the extended difference matrix $\boldsymbol{D}_+$, (2.4). We choose this extended transformation matrix to obtain an extended variance-covariance matrix of random effects that is a direct extension of $\boldsymbol{G}$, the variance-covariance matrix of the random effects in the fit.
Then, following Gilmour et al. (2004), the new predicted values are:

$$\hat{\boldsymbol{y}}_p = \boldsymbol{X}_p\hat{\hat{\boldsymbol{\beta}}} + \boldsymbol{Z}_{(p)}\hat{\hat{\boldsymbol{\alpha}}}, \tag{2.14}$$

with $\boldsymbol{Z}_{(p)} = \boldsymbol{Z}_1 + \boldsymbol{Z}_2\boldsymbol{G}_{po}\boldsymbol{G}^{-1}$ and $\hat{\hat{\boldsymbol{\beta}}}$ and $\hat{\hat{\boldsymbol{\alpha}}}$ the BLUE and BLUP, respectively, estimated from the observed data

$$\begin{bmatrix} \hat{\hat{\boldsymbol{\beta}}} \\ \hat{\hat{\boldsymbol{\alpha}}} \end{bmatrix} = \boldsymbol{Q}^{-1}\begin{bmatrix} \boldsymbol{X}' \\ \boldsymbol{Z}' \end{bmatrix}\boldsymbol{y}, \tag{2.15}$$

where $\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z} + \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{G}^{-1} \end{bmatrix}$. It follows that the predicted random effects vector for $\boldsymbol{\alpha}_p$ is $\hat{\boldsymbol{\alpha}}_p = \boldsymbol{G}_{po}\boldsymbol{G}^{-1}\hat{\hat{\boldsymbol{\alpha}}}$. We use the double hat symbol ($\hat{\ }$) here to remark that the procedure is a two-stage approach, first fit the data and then predict.
Since, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, we have

$$\begin{aligned} \begin{bmatrix} \hat{\hat{\beta}} - \beta \\ \hat{\hat{\boldsymbol{\alpha}}} - \boldsymbol{\alpha} \end{bmatrix} &= \boldsymbol{Q}^{-1}\begin{bmatrix} \boldsymbol{X}' \\ \boldsymbol{Z}' \end{bmatrix}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}) - \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} \\ &= \boldsymbol{Q}^{-1}\begin{bmatrix} \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & -\sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{G}^{-1} \end{bmatrix}\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} + \boldsymbol{Q}^{-1}\begin{bmatrix} \boldsymbol{X}' \\ \boldsymbol{Z}' \end{bmatrix}\boldsymbol{\epsilon}. \end{aligned}$$

It follows that,

$$\begin{aligned}\text{Var}\begin{bmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\end{bmatrix} &= \boldsymbol{Q}^{-1}\begin{bmatrix}\boldsymbol{O} & \boldsymbol{O}\\ \boldsymbol{O} & \sigma_{\boldsymbol{\epsilon}}^4\boldsymbol{G}^{-1}\end{bmatrix}\boldsymbol{Q}^{-1}+\sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{Q}^{-1}\begin{bmatrix}\boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z}\\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z}\end{bmatrix}\boldsymbol{Q}^{-1}\\ &= \sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{Q}^{-1}.\end{aligned}$$

Therefore, the **confidence interval** for a desired confidence level α is:

$$\hat{\boldsymbol{y}}_p \pm Z_{\alpha/2}\sqrt{\sigma_{\boldsymbol{\epsilon}}^2\ \text{diag}\left(\begin{bmatrix}\boldsymbol{X}_p \mid \boldsymbol{Z}_{(p)}\end{bmatrix}\boldsymbol{Q}^{-1}\begin{bmatrix}\boldsymbol{X}_p'\\ \boldsymbol{Z}_{(p)}'\end{bmatrix}\right)},$$

with $\boldsymbol{Q}$ as in (2.15) and the variance components estimated through restricted maximum likelihood procedure (REML) (see Patterson and Thompson 1971). We denote this method as "*mixed model approach*".

Prediction based on the conditional distribution of $\boldsymbol{y}_p|\boldsymbol{y}$

The previous method can be seen from the point of view of conditional distributions, which allow us to compute prediction intervals. We rewrite (2.13) as

$$\boldsymbol{y}_+ = \begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{y}_p\end{bmatrix} \sim \mathcal{N}(\boldsymbol{X}_+\boldsymbol{\beta}, \tilde{\boldsymbol{V}}_+), \tag{2.16}$$

with $\text{Var}[\boldsymbol{y}_+] = \tilde{\boldsymbol{V}}_+ = \boldsymbol{Z}_+\boldsymbol{G}_+\boldsymbol{Z}_+' + \sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I}_+$. The mixed model formulation connects the new values $\boldsymbol{y}_p$ to the observed vector $\boldsymbol{y}$ through a joint normal distribution. It appears, therefore, natural to predict $\boldsymbol{y}_p$ given $\boldsymbol{y}$ based on the conditional model resulting from (2.16), that is

$$\boldsymbol{y}_p|\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}_p\boldsymbol{\beta} + \tilde{\boldsymbol{V}}_{po}\tilde{\boldsymbol{V}}_{oo}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}), \tilde{\boldsymbol{V}}_{pp} - \tilde{\boldsymbol{V}}_{po}\tilde{\boldsymbol{V}}_{oo}^{-1}\tilde{\boldsymbol{V}}_{op}),$$

where $\tilde{\boldsymbol{V}}_{oo}$, $\tilde{\boldsymbol{V}}_{op}$ and $\tilde{\boldsymbol{V}}_{pp}$ are the submatrices of matrix $\tilde{\boldsymbol{V}}_+$ matching to $\boldsymbol{y}$ and $\boldsymbol{y}_p$. The mean value results through

$$\begin{aligned}\mathbb{E}[\boldsymbol{y}_p|\boldsymbol{y}] &= \boldsymbol{X}_p\hat{\hat{\boldsymbol{\beta}}} + \tilde{\boldsymbol{V}}_{po}\tilde{\boldsymbol{V}}_{oo}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})\\ &= \boldsymbol{X}_p\hat{\hat{\boldsymbol{\beta}}} + (\boldsymbol{Z}_1\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{Z}\boldsymbol{G}_{po}\boldsymbol{Z}')(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I})^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})\\ &= \boldsymbol{X}_p\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_1\hat{\boldsymbol{\alpha}} + \boldsymbol{Z}_2\boldsymbol{G}_{po}\boldsymbol{G}^{-1}\hat{\boldsymbol{\alpha}},\end{aligned}$$

which equals (2.14). Note that the first term in the equation above is the result of plugging $\boldsymbol{X}_p$ into the regression equation, and represents the adjustment to this prediction

based on the covariance between $\boldsymbol{y}$ and $\boldsymbol{y}_p$. The conditional variance of $\boldsymbol{y}_p|\boldsymbol{y}$ gives the prediction error. This follows since

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{y},\boldsymbol{y}_p}[(\hat{\boldsymbol{y}}_p - \boldsymbol{y}_p)^2] &= \mathbb{E}_{\boldsymbol{y}}[\mathbb{E}_{\boldsymbol{y}_p}[(\hat{\boldsymbol{y}}_p - \boldsymbol{y}_p)^2|\boldsymbol{y}]] \\
&= \mathbb{E}_{\boldsymbol{y}}[\mathbb{E}_{\boldsymbol{y}_p}[(\boldsymbol{y}_p - \mathbb{E}[\boldsymbol{y}_p|\boldsymbol{y}])^2|\boldsymbol{y}]] \\
&= \mathbb{E}_{\boldsymbol{y}}[\mathrm{Var}[\boldsymbol{y}_p|\boldsymbol{y}]] \\
&= \mathrm{Var}[\boldsymbol{y}_p|\boldsymbol{y}],
\end{aligned}$$

where the latter equality holds since in the Normal model the conditional variance does not depend on the value of the variable we condition on. With these results we can construct the **prediction interval** for a desired confidence level α:

$$\hat{\boldsymbol{y}}_p \pm Z_{\alpha/2}\sqrt{\mathrm{Var}[\boldsymbol{y}_p|\boldsymbol{y}]},$$

where observed values are considered as fixed. Notice that, as we have mentioned before, this approach allow us to work out the posterior predictive distribution $\boldsymbol{y}_p|\boldsymbol{y}$ as a Gaussian distribution and, therefore, compute the prediction intervals. While we can not compute prediction intervals with the standard methodology described in Section 2.2.1 unless we link it with Gaussian processes as we do in Section 2.2.3.

2.2.2 One-stage approach

The previous methods are a two-stage procedure. First, the model is fitted to the available data and second, based on the fitted model we predict the new observations. As mentioned previously, this approach imposes constraints on the reparametrization used to obtain the smooth mixed model, since the variance-covariance matrix of the extended model (the model that includes out-of-sample prediction) needs to be an extension of the variance-covariance matrix of the original model. Now we propose an alternative approach which can be used with any reparametrization and yields the same results as the two step approach. This approach relates the above results to the method presented in Section 2.1, we will call it "*extended mixed model approach*", since we include $\boldsymbol{y}_p$ in the model but with infinite variance (zero weight). In this case we consider the model

$$\boldsymbol{y}_+ = \boldsymbol{X}_+\boldsymbol{\beta} + \boldsymbol{Z}_+\boldsymbol{\alpha}_+ + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\alpha}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}_+), \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+), \tag{2.17}$$

where, for i.i.d. errors, $\boldsymbol{R}_+$ is a diagonal weight matrix of dimension $n_+ \times n_+$, with $\sigma^2_{\boldsymbol{\epsilon}}$ entries if the data is observed, i.e. for $\boldsymbol{y}$, and infinity if the data is considered to be predicted, i.e. for $\boldsymbol{y}_p$. The quantity infinity expresses that we do not have any information

about the data $\boldsymbol{y}_p$. Its estimation is done using the extended mixed model equations of Henderson (1975):

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\tilde{+}} &= (\boldsymbol{X}'_+\boldsymbol{V}_+^-\boldsymbol{X}_+)^{-1}\boldsymbol{X}'_+\boldsymbol{V}_+^-\boldsymbol{y}_+, \\ \hat{\boldsymbol{\alpha}}_+ &= \boldsymbol{G}_+\boldsymbol{Z}'_+\boldsymbol{V}_+^-(\boldsymbol{y}_+ - \boldsymbol{X}_+\hat{\boldsymbol{\beta}}), \end{aligned} \tag{2.18}$$

where $\boldsymbol{y}_+ = (\boldsymbol{y}', \boldsymbol{y}'_p)'$ as in (2.1), $\boldsymbol{V}_+ = \boldsymbol{Z}_+\boldsymbol{G}_+\boldsymbol{Z}'_+ + \boldsymbol{R}_+$ and by Theorem 18.2.8 given in Harville (2000) $\boldsymbol{V}_+^- = \boldsymbol{R}_+^{-1} - \boldsymbol{R}_+^{-1}\boldsymbol{Z}_+(\boldsymbol{G}_+^{-1} + \boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+)^{-1}\boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}$. Finally $\boldsymbol{Z}_+ = \boldsymbol{B}_+\boldsymbol{\Omega}_{+_r}$, with $\boldsymbol{\Omega}_{+_r}$ any transformation such that

$$\boldsymbol{B}_+\left[\boldsymbol{\Omega}_{+_f} \mid \boldsymbol{\Omega}_{+_r}\right] = [\boldsymbol{X}_+ \mid \boldsymbol{Z}_+] \text{ and } \left[\boldsymbol{\Omega}_{+_f} \mid \boldsymbol{\Omega}_{+_r}\right]'\boldsymbol{\theta}_+ = \left[\boldsymbol{\beta}'_{\tilde{+}} \mid \boldsymbol{\alpha}'_+\right]',$$

and $[\boldsymbol{X}_+ \mid \boldsymbol{Z}_+]$ of full rank. The subscript of $\boldsymbol{\beta}_{\tilde{+}}$ is $(\tilde{+})$ and not $(+)$ to indicate that the fixed effects in the extended model (2.17) are not the same as the fixed effects obtained in the fit, however both fixed effects have the same dimension. Notice that to compute the fixed and random effects we do not need $\boldsymbol{R}_+$, we need its inverse, $\boldsymbol{R}_+^{-1}$, with $\frac{1}{\sigma^2_\epsilon}$ entries if the data is observed and 0 if the data is considered to be predicted.

As we mentioned earlier, the above extension of the missing value approach to the mixed model framework fits and predicts simultaneously, while the approach of Gilmour et al. (2004) is a two-stage method. In order to know the relationship between the two methods, we need to know the relationship between the covariance matrix of the random effects that gives the fit, and the extended covariance matrix. This is shown in the following theorem.

Theorem 2.2. *Given model in (1.1) with penalty based on differences between adjacent coefficients, the fit and the prediction of new observations given by extended mixed model approach and mixed model approach are the same if the transformation matrix in (2.18) is the direct extension of the original transformation,*

$$\boldsymbol{\Omega}_{+_r} = \begin{bmatrix} \boldsymbol{\Omega}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Omega}_{p_r} \end{bmatrix}, \tag{2.19}$$

where $\boldsymbol{\Omega}_r = \boldsymbol{U}_r\boldsymbol{\Sigma}^{-1/2}$, based on the SVD of $\boldsymbol{D}'\boldsymbol{D} = \boldsymbol{U}\tilde{\boldsymbol{\Sigma}}\boldsymbol{U}'$, is the transformation matrix for the random component used for the observed data and $\boldsymbol{\Omega}_{p_r} = \boldsymbol{D}_2^{-1}$ is the transformation matrix for the random component of the predicted values, with $\boldsymbol{D}$ and $\boldsymbol{D}_2$ blocks of the extended difference matrix $\boldsymbol{D}_+$, (2.4).

Under the previous hypothesis the variance components ($\sigma^2_{\boldsymbol{\alpha}}$, σ^2_ϵ) that maximize the

REML, l, and the REML corresponding to the extended weighted model, l_+, are equal,

$$l(\sigma^2_{\boldsymbol{\epsilon}}, \sigma^2_{\boldsymbol{\alpha}}) = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right), \tag{2.20}$$

$$l_+(\sigma^2_{\boldsymbol{\epsilon}}, \sigma^2_{\boldsymbol{\alpha}}) = -\frac{1}{2}\log|\boldsymbol{V}_+| - \frac{1}{2}\log|\boldsymbol{X}'_+\boldsymbol{V}^-_+\boldsymbol{X}_+| - \frac{1}{2}\left(\boldsymbol{y}_+ - \boldsymbol{X}_+\boldsymbol{\beta}\right)'\boldsymbol{V}^-_+\left(\boldsymbol{y}_+ - \boldsymbol{X}_+\boldsymbol{\beta}\right). \tag{2.21}$$

Equivalent results would be obtained using maximum likelihood (ML).

Proof. Since with the transformation matrix (2.19) the extended fixed and random parts are the same in both methods, we just need to show that the fixed and random effects are equal in both methods.

Let us compute the covariance matrix $\boldsymbol{G}_+$ of the augmented random effects $\boldsymbol{\alpha}_+$, (2.18):

$$\boldsymbol{G}_+ = \sigma^2_{\boldsymbol{\alpha}}(\boldsymbol{\Omega}'_{+_r}\boldsymbol{D}'_+\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1} = \begin{bmatrix} \boldsymbol{G} & \boldsymbol{G}_{op} \\ \boldsymbol{G}_{po} & \boldsymbol{G}_{pp} \end{bmatrix},$$

$\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r}$ is a squared matrix, so the inverse of $((\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})'\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1}$ is

$$((\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})'\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1} = (\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1}(\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{'-1} = (\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1}(\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1'}.$$

Using Lemma 8.5.4 of Harville (2000), we have that:

$$(\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1} = \begin{bmatrix} \boldsymbol{D}\boldsymbol{\Omega}_r & \boldsymbol{0} \\ \boldsymbol{D}_1\boldsymbol{\Omega}_r & \boldsymbol{D}_2\boldsymbol{\Omega}_{p_r} \end{bmatrix}^{-1} = \begin{bmatrix} (\boldsymbol{D}\boldsymbol{\Omega}_r)^{-1} & \boldsymbol{0} \\ -(\boldsymbol{D}_2\boldsymbol{\Omega}_{p_r})^{-1}\boldsymbol{D}_1\boldsymbol{\Omega}_r(\boldsymbol{D}\boldsymbol{\Omega}_r)^{-1} & (\boldsymbol{D}_2\boldsymbol{\Omega}_{p_r})^{-1} \end{bmatrix}.$$

Therefore,

$$\boldsymbol{G} = \sigma^2_{\alpha}\boldsymbol{I}, \quad \boldsymbol{G}_{op} = \sigma^2_{\alpha}(-\boldsymbol{\Omega}'_r\boldsymbol{D}'_1), \quad \boldsymbol{G}_{po} = \sigma^2_{\alpha}(-\boldsymbol{D}_1\boldsymbol{\Omega}_r), \quad \boldsymbol{G}_{pp} = \sigma^2_{\alpha}(\boldsymbol{I} + \boldsymbol{D}_1\boldsymbol{\Omega}_r\boldsymbol{\Omega}'_r\boldsymbol{D}'_1).$$

Notice that its inverse is:

$$\boldsymbol{G}^{-1}_+ = \begin{bmatrix} \boldsymbol{G}^{oo} & \boldsymbol{G}^{op} \\ \boldsymbol{G}^{po} & \boldsymbol{G}^{pp} \end{bmatrix} = \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\begin{bmatrix} \boldsymbol{I} + \boldsymbol{\Omega}'_r\boldsymbol{D}'_1\boldsymbol{D}_1\boldsymbol{\Omega}_r & -\boldsymbol{\Omega}'_r\boldsymbol{D}'_1 \\ \boldsymbol{D}_1\boldsymbol{\Omega}_r & \boldsymbol{I} \end{bmatrix}.$$

Now that we know $\boldsymbol{G}_+$, we just need to compute $\boldsymbol{V}^-_+$ to know the expression of the extended fixed effects. Since $\boldsymbol{V}^-_+ = \boldsymbol{R}^{-1}_+ - \boldsymbol{R}^{-1}_+\boldsymbol{Z}_+(\boldsymbol{G}^{-1}_+ + \boldsymbol{Z}'_+\boldsymbol{R}^{-1}_+\boldsymbol{Z}_+)^{-1}\boldsymbol{Z}'_+\boldsymbol{R}^{-1}_+$, we have that,

$$\boldsymbol{G}^{-1}_+ + \boldsymbol{Z}'_+\boldsymbol{R}^{-1}_+\boldsymbol{Z}_+ = \begin{bmatrix} \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}(\boldsymbol{I} + \boldsymbol{\Omega}'_r\boldsymbol{D}'_1\boldsymbol{D}_1\boldsymbol{\Omega}_r) + \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}(\boldsymbol{B}\boldsymbol{\Omega}_r)'\boldsymbol{B}\boldsymbol{\Omega}_r & \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{\Omega}'_r\boldsymbol{D}'_1 \\ \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{D}_1\boldsymbol{\Omega}_r & \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{K}_1 & \boldsymbol{K}_2 \\ \boldsymbol{K}_3 & \boldsymbol{K}_4 \end{bmatrix},$$

and that,

$$\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+ = \begin{bmatrix} \frac{1}{\sigma_\epsilon^2}\boldsymbol{B\Omega}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}.$$

Defining $\left(\boldsymbol{G}_+^{-1} + \boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+\right)^{-1} = \begin{bmatrix} \boldsymbol{J}_1 & \boldsymbol{J}_2 \\ \boldsymbol{J}_3 & \boldsymbol{J}_4 \end{bmatrix}$, it follows that:

$$\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+\left(\boldsymbol{G}_+^{-1} + \boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+\right)^{-1}\boldsymbol{Z}'_+\boldsymbol{R}_+^{-1} = \begin{bmatrix} \frac{1}{\sigma_\epsilon^4}\boldsymbol{B\Omega}_r\boldsymbol{J}_1(\boldsymbol{B\Omega}_r)' & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}.$$

Hence, we just need to know $\boldsymbol{J}_1$. Applying Theorem 8.5.11 given in Harville (2000):

$$\begin{aligned} \boldsymbol{J}_1^{-1} &= \boldsymbol{K}_1 - \boldsymbol{K}_2\boldsymbol{K}_4^{-1}\boldsymbol{K}_3 \\ &= \boldsymbol{K}_1 - \frac{1}{\sigma_{\boldsymbol{\alpha}}^2}\boldsymbol{\Omega}'_r\boldsymbol{D}'_1(\sigma_{\boldsymbol{\alpha}}^2\boldsymbol{I})\frac{1}{\sigma_{\boldsymbol{\alpha}}^2}\boldsymbol{D}_1\boldsymbol{\Omega}_r \\ &= \frac{1}{\sigma_{\boldsymbol{\alpha}}^2}(\boldsymbol{I} + \boldsymbol{\Omega}_r^{'}\boldsymbol{D}_1^{'}\boldsymbol{D}_1\boldsymbol{\Omega}_r) + \frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}(\boldsymbol{B\Omega}_r)'\boldsymbol{B\Omega}_r - \frac{1}{\sigma_{\boldsymbol{\alpha}}^2}\boldsymbol{\Omega}'_r\boldsymbol{D}'_1\boldsymbol{D}_1\boldsymbol{\Omega}_r \\ &= \frac{1}{\sigma_{\boldsymbol{\alpha}}^2}\boldsymbol{I} + \frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}(\boldsymbol{B\Omega}_r)'\boldsymbol{B\Omega}_r, \end{aligned}$$

and, applying Theorem 18.2.8, given in Harville (2000) to compute $\boldsymbol{J}_1$:

$$\boldsymbol{J}_1 = \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{I} - (\sigma_{\boldsymbol{\alpha}}^2)^2(\boldsymbol{B\Omega}_r)'(\sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I} + \boldsymbol{B\Omega}_r\sigma_{\boldsymbol{\alpha}}^2\boldsymbol{I}(\boldsymbol{B\Omega}_r)')^{-1}\boldsymbol{B\Omega}_r.$$

Therefore:

$$\begin{aligned} \boldsymbol{V}_+^- &= \boldsymbol{R}_+^{-1} - \boldsymbol{R}_+^{-1}\boldsymbol{Z}_+\left(\boldsymbol{G}_+^{-1} + \boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+\right)^{-1}\boldsymbol{Z}'_+\boldsymbol{R}_+^{-1} \\ &= \begin{bmatrix} \frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}\boldsymbol{I} - \frac{1}{\sigma_{\boldsymbol{\epsilon}}^4}\boldsymbol{B\Omega}_r\left[\sigma_{\boldsymbol{\alpha}}^2\boldsymbol{I} - \sigma_{\boldsymbol{\alpha}}^4(\boldsymbol{B\Omega}_r)'(\sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I} + \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{B\Omega}_r(\boldsymbol{B\Omega}_r)')^{-1}\boldsymbol{B\Omega}_r\right](\boldsymbol{B\Omega}_r)' & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{V}_{+11}^- & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}. \end{aligned}$$

Moreover, as $\boldsymbol{V} = \sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I} + \boldsymbol{ZGZ}'$, with $\boldsymbol{G} = \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{I}$:

$$\boldsymbol{V}^{-1} = \frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}\boldsymbol{I} - \frac{1}{\sigma_{\boldsymbol{\epsilon}}^4}\boldsymbol{B\Omega}_r\left(\frac{1}{\sigma_{\boldsymbol{\alpha}}^2}\boldsymbol{I} + \frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}(\boldsymbol{B\Omega}_r)'\boldsymbol{B\Omega}_r\right)^{-1}(\boldsymbol{B\Omega}_r)'.$$

By Theorem 18.2.8 given in Harville (2000),

$$\left(\frac{1}{\sigma_{\boldsymbol{\alpha}}^2}\boldsymbol{I} + \frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}(\boldsymbol{B\Omega}_r)'\boldsymbol{B\Omega}_r\right)^{-1} = \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{I} - \sigma_{\boldsymbol{\alpha}}^4(\boldsymbol{B\Omega}_r)'\left(\sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I} + \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{B\Omega}_r(\boldsymbol{B\Omega}_r)'\right)^{-1}\boldsymbol{B\Omega}_r$$

i.e., $\boldsymbol{V}^{-1} = \boldsymbol{V}_{+11}^{-}$.

As we have proved that $\boldsymbol{V}_{+}^{-} = \begin{bmatrix} \boldsymbol{V}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}$ it is straightforward to show that $\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}}$.

Moreover, by the extended mixed model approach we have that,

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_{+} &= \boldsymbol{G}_{+}\boldsymbol{Z}'_{+}\hat{\hat{\boldsymbol{V}}}_{+}^{-}(\boldsymbol{y}_{+} - \boldsymbol{X}_{+}\hat{\boldsymbol{\beta}}) \\ &\underset{\hat{\boldsymbol{\beta}}=\hat{\hat{\boldsymbol{\beta}}}}{=} \begin{bmatrix} (\boldsymbol{B}\boldsymbol{\Omega}_r)'\hat{\boldsymbol{V}}^{-1} & \boldsymbol{O} \\ -\boldsymbol{D}_1\boldsymbol{\Omega}_r(\boldsymbol{B}\boldsymbol{\Omega}_r)'\hat{\boldsymbol{V}}^{-1} & \boldsymbol{O} \end{bmatrix} (\boldsymbol{y}_{+} - \boldsymbol{X}_{+}\hat{\hat{\boldsymbol{\beta}}}) \\ &= \begin{bmatrix} \boldsymbol{G}\boldsymbol{Z}'\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}) \\ \boldsymbol{G}_{po}\boldsymbol{G}^{-1}\boldsymbol{G}\boldsymbol{Z}'\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}) \end{bmatrix} \\ &= \begin{bmatrix} \hat{\hat{\boldsymbol{\alpha}}} \\ \boldsymbol{G}_{po}\boldsymbol{G}^{-1}\hat{\hat{\boldsymbol{\alpha}}} \end{bmatrix}. \end{aligned}$$

As we wanted to show solutions given by extended mixed model approach and mixed model approach are the same.

Let us prove that the variance components ($\sigma^2_{\boldsymbol{\epsilon}}$, $\sigma^2_{\boldsymbol{\alpha}}$) that maximize the approximate restricted maximum likelihoods (2.20) and (2.21) are equal. Consider the parts of both expressions as follows:

$$l(\sigma^2_{\boldsymbol{\epsilon}}, \sigma^2_{\boldsymbol{\alpha}}) = \underbrace{-\frac{1}{2}\log|\boldsymbol{V}|}_{\text{Part I}} \underbrace{-\frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|}_{\text{Part II}} \underbrace{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}_{\text{Part III}},$$

$$l_{+}(\sigma^2_{\boldsymbol{\epsilon}}, \sigma^2_{\boldsymbol{\alpha}}) = \underbrace{-\frac{1}{2}\log|\boldsymbol{V}_{+}|}_{\text{Part I}} \underbrace{-\frac{1}{2}\log|\boldsymbol{X}'_{+}\boldsymbol{V}_{+}^{-}\boldsymbol{X}_{+}|}_{\text{Part II}} \underbrace{-\frac{1}{2}(\boldsymbol{y}_{+} - \boldsymbol{X}_{+}\boldsymbol{\beta})'\boldsymbol{V}_{+}^{-}(\boldsymbol{y}_{+} - \boldsymbol{X}_{+}\boldsymbol{\beta})}_{\text{Part III}},$$

since $\boldsymbol{V}_{+}^{-} = \begin{bmatrix} \boldsymbol{V}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}$, it is straightforward to prove that Part II and Part III of both restricted maximum likelihoods are equal. As $\boldsymbol{V}_{+} \neq \boldsymbol{V}$, Part I of (2.20) and (2.21) are not equal, but its derivatives with respect to the parameters ($\sigma^2_{\boldsymbol{\epsilon}}$, $\sigma^2_{\boldsymbol{\alpha}}$) are equal:

Derivatives of Part I with respect to $\sigma^2_{\boldsymbol{\epsilon}}$:

$$\frac{\partial\left(\frac{1}{2}\log|\boldsymbol{V}|\right)}{\partial\sigma^2_{\boldsymbol{\epsilon}}} = \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\right)$$

and

$$\frac{\partial\left(\frac{1}{2}\log|\boldsymbol{V}_+|\right)}{\partial\sigma^2_{\boldsymbol{\epsilon}}} = \frac{1}{2}\text{trace}\left(\boldsymbol{V}_+^-\frac{\partial\boldsymbol{R}_+}{\partial\sigma^2_{\boldsymbol{\epsilon}}}\right) = \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\right),$$

Derivatives of Part I with respect to $\sigma^2_{\boldsymbol{\alpha}}$:

$$\frac{\partial\left(\frac{1}{2}\log|\boldsymbol{V}|\right)}{\partial\sigma^2_{\boldsymbol{\alpha}}} = \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'\right)$$

and

$$\begin{aligned}
\frac{\partial\left(\frac{1}{2}\log|\boldsymbol{V}_+|\right)}{\partial\sigma^2_{\boldsymbol{\alpha}}} &= \frac{1}{2}\text{trace}\left(\boldsymbol{V}_+^-\boldsymbol{Z}_+\frac{\partial\boldsymbol{G}_+}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'_+\right) \\
&= \frac{1}{2}\text{trace}\left(\begin{bmatrix}\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' & \boldsymbol{V}^{-1}\boldsymbol{Z}\left(\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'_1 + \frac{\partial\boldsymbol{G}_{op}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'_2\right)\\ \boldsymbol{O} & \boldsymbol{O}\end{bmatrix}\right) \\
&= \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'\right).
\end{aligned}$$

■

Since the fitted and predicted values in both approaches do not depend on the used transformations, the previous theorem is stating a stronger result. The approaches always give the same solution, regardless of the used transformations. We have stated the theorem for a particular transformation because otherwise the fixed and random parts and fixed and random effects of both methods do not have to be the same and we could not have established the relationship between both methods.

The last statement of the previous theorem means that the variance parameters used to predict are the same as the ones used for estimating the original fit. In other words prediction within and out-of-sample can not only be done simultaneously, but also the optimal smoothing parameter is the same.

2.2.3 Prediction in the context of penalized Gaussian process regression

As it is known, one of the attractive features of penalized regression is its link to stochastic processes. The title of this section is inspired by the work in Yi et al. (2011). Where they use the penalized Gaussian process regression to provide an alternative solution to the Gaussian process regression variable selection problem, since when dimension of the data is high, it suffers from large variance of parameter estimation and high predictive

errors. They apply several penalized methods to a Gaussian process model, including Ridge, LASSO, Bridge, SCAD and adaptive LASSO penalties.

We also use Gaussian processes but not on the curve; we make a representation of the curve in terms of bases and coefficients, and we have a Gaussian process on the coefficients. Our proposal to predict new values, in the context of Gaussian process smoothing, is to use a model based on Gaussian process prior and a P-spline covariance matrix to fit non linear data. That is, we will harness the flexibility of the Gaussian process and the choice of a suitable covariance matrix to model any nonlinear model nonparametrically. In addition, the prediction is quite straightforward due to the properties of Gaussian processes.

Prediction with Gaussian processes has a long history in at least three literatures: mathematical geology (where the approach is called 'kriging', see Cressie 1993), neural networks (Poggio and Girosi 1990 and Girosi et al. 1995) and global optimization in the analysis of computer experiments (e.g. Sacks et al. 1989).

Building on the previous approaches, we predict new values by proposing the penalized regression framework to a Gaussian process model. In the context of model (1.1), we can assume that the stochastic behaviour of the random vector $\boldsymbol{y}$ depends on the observed covariate and a latent process $\boldsymbol{s}$, according to a linear mixed model,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{s} + \boldsymbol{\epsilon},$$

with $\boldsymbol{X}\boldsymbol{\beta}$ the trend and $\boldsymbol{\epsilon}$ an independent Gaussian process with zero mean and variance $\sigma^2_{\boldsymbol{\epsilon}}$, modelling the measurement error, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I})$. Random effects are assumed to account for variability and represented in terms of basis functions $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{\alpha}$, with $\boldsymbol{B}$ any basis. Imposing a prior structure on $\boldsymbol{\alpha}$ through $\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{P}^-)$, with $\boldsymbol{P}^-$ the covariance matrix of the vector of coefficients, we have the Gaussian process

$$\boldsymbol{s} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{ss}),$$

with $\boldsymbol{\Sigma}_{ss} = \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{B}\boldsymbol{P}^-\boldsymbol{B}'$. Independence of $\boldsymbol{s}$ and $\boldsymbol{\epsilon}$ implies that elements of $\boldsymbol{y}$ are independent and normally distributed conditionally on $\boldsymbol{X}$ and $\boldsymbol{s}$. Then, the marginal distribution of the process $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{yy}}),$$

with $\boldsymbol{\Sigma}_{\boldsymbol{yy}} = \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I} + \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{B}\boldsymbol{P}^-\boldsymbol{B}'$.
Assuming that the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{yy}}$ is known, the maximum likelihood estimator

of the trend parameter vector $\boldsymbol{\beta}$ is

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{y}. \tag{2.22}$$

As is well known from Normal distribution theory, the conditional normal distribution of $\boldsymbol{s}|\boldsymbol{y}$ is $\mathcal{N}(\mathbb{E}[\boldsymbol{s}|\boldsymbol{y}], \boldsymbol{\Sigma}_{s|y})$, with:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{s}|\boldsymbol{y}] &= \mathbb{E}[\boldsymbol{s}] + \boldsymbol{\Sigma}_{sy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}]) \\
&= \boldsymbol{0} + \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}'(\sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I} + \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}')^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \boldsymbol{B}(\lambda\boldsymbol{P} + \boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),
\end{aligned}$$

since $\boldsymbol{\Sigma}_{sy} = \mathbb{E}[\boldsymbol{s}\boldsymbol{s}] = \boldsymbol{\Sigma}_{ss} = \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}'$, and

$$\begin{aligned}
\boldsymbol{\Sigma}_{s|y} &= \boldsymbol{\Sigma}_{ss} - \boldsymbol{\Sigma}_{sy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{ys} \\
&= \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}' - \sigma_{\boldsymbol{\alpha}}^4\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}'(\sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I} + \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}')^{-1}\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}' \\
&= \sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{B}\left[\lambda\boldsymbol{P} + \boldsymbol{B}'\boldsymbol{B}\right]^{-1}\boldsymbol{B}'.
\end{aligned}$$

Therefore, the fit is:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{s}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{B}(\lambda\boldsymbol{P} + \boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{2.23}$$

Let $\boldsymbol{x}_p$ be a vector of n_p unobserved values of the process with

$$\boldsymbol{y}_p = \boldsymbol{X}_p\boldsymbol{\beta} + \boldsymbol{s}_p + \boldsymbol{\epsilon}_p,$$

where $\boldsymbol{s}_p \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{\Sigma}_{s_p s_p})$, $\boldsymbol{\epsilon}_p \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I})$. Therefore, the joint distribution of observed and unobserved values is given by:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y}_p \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{X}_p\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yy_p} \\ \boldsymbol{\Sigma}'_{yy_p} & \boldsymbol{\Sigma}_{y_p y_p} \end{bmatrix}\right),$$

where $\boldsymbol{\Sigma}_{yy} = \sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I} + \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{\Sigma}_{ss}$ and $\boldsymbol{\Sigma}_{yy_p} = \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{\Sigma}_{ss_p}$.

Pollice and Bilancia (2002) showed that the minimum variance predictor of $\boldsymbol{y}_p$ conditional on values of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{yy}$, is given by

$$\mathbb{E}[\boldsymbol{y}_p|\boldsymbol{y}] = \boldsymbol{X}_p\boldsymbol{\beta} + \boldsymbol{\Sigma}'_{yy_p}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Therefore, in order to calculate the predicted values we need to compute $\boldsymbol{\Sigma}'_{\boldsymbol{yy}_p}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{yy}}$. Imposing a prior structure on $\boldsymbol{\alpha}_+$ through $\boldsymbol{\alpha}_+ \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{P}^-_+)$, where $\boldsymbol{P}^-_+$ is the covariance of the extended vector of coefficients:

$$\boldsymbol{P}^-_+ = \begin{bmatrix} \boldsymbol{P}^1 & \boldsymbol{P}^2 \\ \boldsymbol{P}^{2'} & \boldsymbol{P}^3 \end{bmatrix},$$

and since the extended basis is $\boldsymbol{B}_+$ is (2.2), we have:

$$\boldsymbol{B}_+\boldsymbol{P}^-_+\boldsymbol{B}'_+ = \begin{bmatrix} \boldsymbol{B}\boldsymbol{P}^1\boldsymbol{B}' & \boldsymbol{B}(\boldsymbol{P}^1\boldsymbol{B}'_1 + \boldsymbol{P}^2\boldsymbol{B}'_2) \\ (\boldsymbol{B}_1\boldsymbol{P}^1 + \boldsymbol{B}_2\boldsymbol{P}^{2'})\boldsymbol{B}' & (\boldsymbol{B}_1\boldsymbol{P}^1 + \boldsymbol{B}_2\boldsymbol{P}^{2'})\boldsymbol{B}'_1 + (\boldsymbol{B}_1\boldsymbol{P}^2 + \boldsymbol{B}_2\boldsymbol{P}^3)\boldsymbol{B}'_2 \end{bmatrix},$$

i.e., $\boldsymbol{\Sigma}_{\boldsymbol{yy}_p} = \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{B}(\boldsymbol{P}^1\boldsymbol{B}'_1 + \boldsymbol{P}^2\boldsymbol{B}'_2)$. Then, applying Theorem 18.2.8 and Lemma 18.2.1 given in Harville (2000), we have that:

$$\begin{aligned} \boldsymbol{\Sigma}'_{\boldsymbol{yy}_p}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{yy}} &= \sigma^2_{\boldsymbol{\alpha}}(\boldsymbol{B}_1\boldsymbol{P}^1\boldsymbol{B}' + \boldsymbol{B}_2\boldsymbol{P}^{2'}\boldsymbol{B}')(\sigma^2_{\boldsymbol{\alpha}}\boldsymbol{B}\boldsymbol{P}^-\boldsymbol{B}' + \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I})^{-1} \\ &= \sigma^{-2}_{\boldsymbol{\epsilon}}\boldsymbol{B}_1\boldsymbol{P}^1\boldsymbol{P}(\sigma^{-2}_{\boldsymbol{\alpha}}\boldsymbol{P} + \sigma^{-2}_{\boldsymbol{\epsilon}}\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}' + \sigma^{-2}_{\boldsymbol{\epsilon}}\boldsymbol{B}_2\boldsymbol{P}^{2'}\boldsymbol{P}(\sigma^{-2}_{\boldsymbol{\alpha}}\boldsymbol{P} + \sigma^{-2}_{\boldsymbol{\epsilon}}\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'. \end{aligned}$$

Therefore, predictions are written as a function of the extended penalty matrix:

$$\hat{\boldsymbol{y}}_p = \boldsymbol{X}_p\hat{\boldsymbol{\beta}} + \begin{bmatrix} \boldsymbol{B}_1 & \boldsymbol{B}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{I} \\ -\boldsymbol{P}^-_3\boldsymbol{P}'_2 \end{bmatrix} (\boldsymbol{P}_1 - \boldsymbol{P}_2\boldsymbol{P}^{-1}_3\boldsymbol{P}'_2)^-\boldsymbol{P}(\boldsymbol{B}'\boldsymbol{B} + \lambda\boldsymbol{P})^{-1}\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{2.24}$$

with $\hat{\boldsymbol{\beta}}$ as in (2.22).

Since, essentially splines correspond to Gaussian processes with a particular choice of covariance function, in the case of penalties based on differences the extended penalty matrix is (2.5), it is straightforward to prove that the solution of the missing value approach (2.11) and the solution assuming that the response is a realization of a Gaussian process (2.24) are equal.

2.3 Applications

In this section, we apply the proposed methods to three real data sets. The first one allows us to show an example of predicting within the framework of additive models and in the case when out-of-sample prediction is needed to the left and right of the interval where the covariate is observed. The second dataset illustrates a classic example where forecasting is needed, when data are collected over time, and the third dataset shows an

example where the errors are correlated.

2.3.1 Prediction of aboveground biomass

All the results presented in the previous sections are obtained in the case of smooth models with a single covariate. However, it is immediate to extend these results to the case of semi-parametric or additive models. In this section, we apply the proposed methodology to such data set. The data set corresponds to an agricultural trial carried out in Spain (Rivas-Martínez et al. 2002) with the aim of evaluating the economic viability of *Populus* trees prior to harvesting. In this context, it is essential to estimate aboveground biomass and obtain accurate predictions using only a minimum set of easily obtainable information i.e., diameter and height. Sánchez-González et al. (2016) proposed the use of smooth additive mixed models for predicting aboveground biomass. Here we analyze data for a single clone of the nine included in the trials. The aim is to estimate the production (measured as aboveground dry biomass) as a function of diameter and height of the tree and give out-of-sample predictions. The observed data consists of 315 observations, for diameter values measured at 1.30 m breast height. This data is illustrated in Figure 2.3.

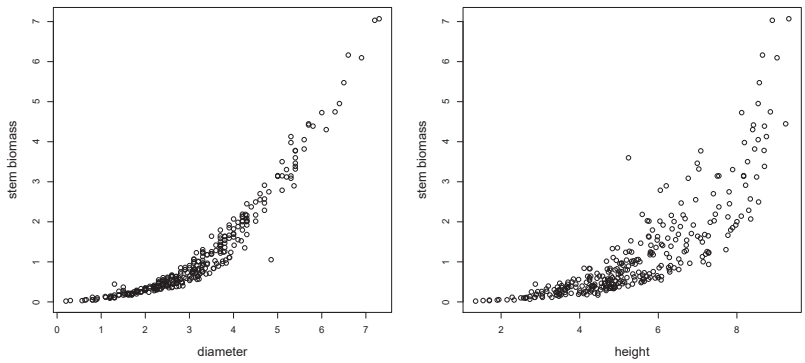


Figure 2.3: Plot of weight versus diameter (left panel) and plot of weight versus height (right panel).

From the plot, it is immediate to notice that the weight is a non-linear function of diameter and height. Therefore we fit the following model:

$$\mathbb{E}[y_i|x_i, y_i] = f(x_i) + f(z_i), \tag{2.25}$$

where $\boldsymbol{x}$ is the diameter and $\boldsymbol{z}$ is the height, i.e. $f(\boldsymbol{x})$ and $f(\boldsymbol{z})$ are the functions that represent the main effects of the diameter and of the height, respectively. The regression matrix is then defined by blocks as $\boldsymbol{B} = [\boldsymbol{B_x} \mid \boldsymbol{B_z}]$ with marginal B-spline bases of degree

three of the covariates diameter and height, $\boldsymbol{B_x}$ and $\boldsymbol{B_z}$, respectively. The penalty matrix associated to model (2.25) has a block-diagonal form: $\boldsymbol{P} = blockdiag(\lambda_{\boldsymbol{x}}\boldsymbol{P_x}, \lambda_{\boldsymbol{z}}\boldsymbol{P_z})$, where $\boldsymbol{P_x}$ and $\boldsymbol{P_z}$ are the marginal second-order difference penalties for diameter and height. We predict weight for 6 new out-of-sample values for diameter and 15 new values for height where 5 are to the left and 10 to the right of the range of the observed height values. Applying the previous methodology we extend the basis and the penalty: $\boldsymbol{B}_+ = \left[\boldsymbol{B}_{\boldsymbol{x}_+} \mid \boldsymbol{B}_{\boldsymbol{z}_+}\right]$, $\boldsymbol{P}_+ = blockdiag(\lambda_{\boldsymbol{x}}\boldsymbol{P}_{\boldsymbol{x}_+}, \lambda_{\boldsymbol{z}}\boldsymbol{P}_{\boldsymbol{z}_+})$, with $\boldsymbol{B}_{\boldsymbol{x}_+} = \begin{bmatrix} \boldsymbol{B_x} & \boldsymbol{O} \\ \boldsymbol{B}_{\boldsymbol{x}_{(1)}} & \boldsymbol{B}_{\boldsymbol{x}_{(2)}} \end{bmatrix}$ and $\boldsymbol{B}_{\boldsymbol{z}_+} = \begin{bmatrix} \boldsymbol{B}_{\boldsymbol{z}_{(0)}} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{B_z} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{B}_{\boldsymbol{z}_{(1)}} & \boldsymbol{B}_{\boldsymbol{z}_{(2)}} \end{bmatrix}$.

Once we have extended the basis and the penalty, it is straightforward to obtain the fit and the prediction applying (2.1). However, in order to obtain confidence and prediction intervals and to avoid identifiability problems (since the column of **1**'s is contained in the space spanned by the columns of $\boldsymbol{B}_{\boldsymbol{x}_+}$ and $\boldsymbol{B}_{\boldsymbol{z}_+}$), we reparameterize the model using the representation of a penalized spline model as a mixed model. Figure 2.4 shows the smooth fitted and predicted trend for diameter and height, the 95% confidence interval (grey line) and the 95% prediction interval (dashed lines). Notice that the prediction is done backward and forward. This is important, since the proposed methodology allows us to obtain the prediction for any range of the independent variable. This could not be done by using methods developed in the series temporal framework.

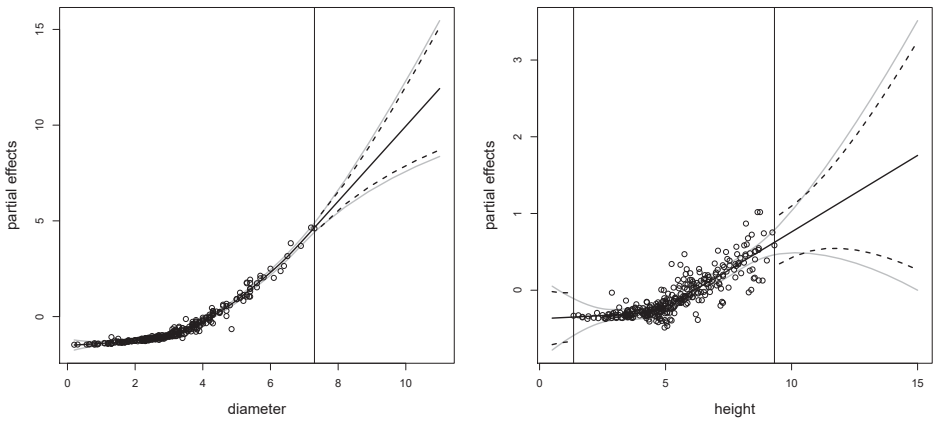


Figure 2.4: Fit, forecast, 95% confidence interval (grey lines) and 95% prediction interval (dashed lines) of the additive smooth term for diameter (left panel) and for height (right panel), result of applying the methodology of a data set on the stem biomass.

The estimation of the covariance parameters was carried out by the REML through

the SOP algorithm, proposed in Rodríguez-Álvarez et al. (2018), i.e. the smoothing parameters are computed as the ratios between variance parameters that are estimated in the mixed models framework.

2.3.2 Forecasting SO_2 concentration levels

Now, we analyze data where out-of-sample prediction is needed, and missing observations are present in the data. We consider measurements on sulphur dioxide (SO_2) concentration levels (in $\mu g/m^3$) over station AT02 from January 1990 to December 2001, Figure 2.5 shows the data set, in which we can see that there are some missing observations between October 1995 and March 1999.

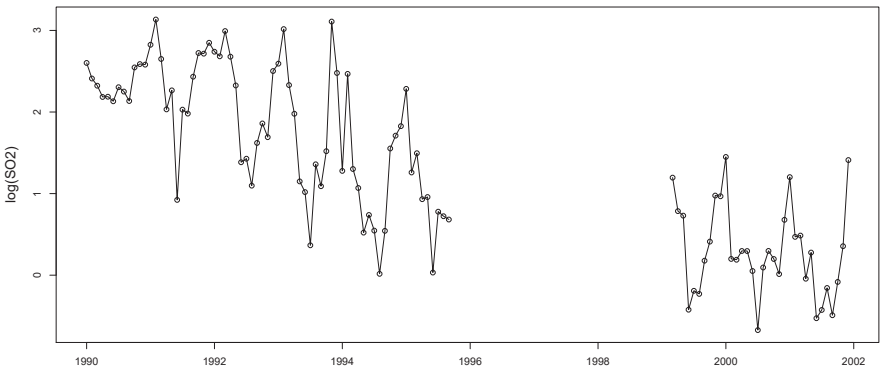


Figure 2.5: Time series plot of $log(SO_2)$ data for station AT02.

The data were collected through the 'European monitoring and evaluation programme' (EMEP) under the Co-operative Programme for Monitoring and Evaluation of the Long-range Transmission of Air Pollutants in Europe (see further information available at `htpp:\\www.emep.int`) and also used in Lee and Durbán (2012). Notice that there is a clear evidence of temporal trends and seasonal effects and that there are some missing observations, there is a big gap between October 1995 and March 1999.
First of all, let us introduce the smooth modulation model based on P-splines suggested by Eilers et al. (2008):

$$\mathbb{E}\left[y_i|\, x_i\right] = f(x_i) + \sum_{j=1}^{J} \{g_j(x_i)\cos(j\omega x_i) + h_j(x_i)\sin(j\omega x_i)\}, \tag{2.26}$$

where $f(\cdot)$ accounts for the smooth trend, $g(\cdot)$ and $h(\cdot)$ are smooth functions that describe the local amplitudes of cosine and sine waves, and $\omega = 2\pi/p$, with p the period (e.g. $p = 12$ for monthly data). The number of harmonics functions, J, required for the

seasonal component is usually taken as 1 or 2 to reduce the number of parameters to be estimated. If $J = 1$ the model (2.26) can be written in matrix form as:

$$\boldsymbol{y} = \breve{\boldsymbol{B}}\breve{\boldsymbol{\theta}} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}), \tag{2.27}$$

where $\boldsymbol{R}$ is the covariance matrix of the error, we work with uncorrelated i.i.d. errors, i.e., $\boldsymbol{R} = \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}$; and $\breve{\boldsymbol{B}}$ is the regression matrix, $\breve{\boldsymbol{B}} = [\boldsymbol{B} \mid \boldsymbol{CB} \mid \boldsymbol{SB}]$, where $\boldsymbol{B}$ is a B-spline basis, $\boldsymbol{C} = \text{diag}\{\cos(\omega \boldsymbol{x})\}$ and $\boldsymbol{S} = \text{diag}\{\sin(\omega \boldsymbol{x})\}$.

If we want to obtain the coefficients $\breve{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \boldsymbol{\theta}_c, \boldsymbol{\theta}_s)$ in (2.27) with the P-splines method we have to minimize the following function of $\breve{\boldsymbol{\theta}}$:

$$S = (\boldsymbol{y} - \breve{\boldsymbol{B}}\breve{\boldsymbol{\theta}})'(\boldsymbol{y} - \breve{\boldsymbol{B}}\breve{\boldsymbol{\theta}}) + \breve{\boldsymbol{\theta}}'\breve{\boldsymbol{P}}\breve{\boldsymbol{\theta}}, \tag{2.28}$$

where $\breve{\boldsymbol{P}} = \breve{\boldsymbol{\lambda}} \otimes \breve{\boldsymbol{D}}'\breve{\boldsymbol{D}}$, with $\breve{\boldsymbol{\lambda}} = diag(\lambda, \lambda_c, \lambda_s)$ and $\breve{\boldsymbol{D}} = \text{blockdiag}(\boldsymbol{D}'_q\boldsymbol{D}'_q, \boldsymbol{I}_2 \otimes \boldsymbol{D}'_{q_{cs}}\boldsymbol{D}_{q_{cs}})$, q and q_{cs} are usually 2 and 1, respectively.

To obtain the fit and the forecast simultaneously, we just have to extend the B-spline basis for the trend and the modulation components $\breve{\boldsymbol{B}}_+ = [\boldsymbol{B}_+ \mid \boldsymbol{C}_+\boldsymbol{B}_+ \mid \boldsymbol{S}_+\boldsymbol{B}_+]$, where $\boldsymbol{C}_+ = \text{diag}(\cos(\omega \boldsymbol{x}_+))$ and $\boldsymbol{S}_+ = \text{diag}(\sin(\omega \boldsymbol{x}_+))$, for the additive modulation blocks $\boldsymbol{C}_+\boldsymbol{B}_+$ and $\boldsymbol{S}_+\boldsymbol{B}_+$, and also consider the following penalty matrix:

$$\breve{\boldsymbol{P}}_+ = \text{blockdiag}(\lambda \boldsymbol{D}'_{+q}\boldsymbol{D}_{+q}, \lambda_c \boldsymbol{D}'_{+q_c}\boldsymbol{D}_{+q_c}, \lambda_s \boldsymbol{D}'_{+q_s}\boldsymbol{D}_{+q_s}).$$

Once $\breve{\boldsymbol{B}}_+$ and $\breve{\boldsymbol{P}}_+$ are computed, $\breve{\boldsymbol{\theta}}_+$ can be easily computed through the formula in (2.1), by using $\breve{\boldsymbol{B}}_+$ instead of $\boldsymbol{B}_+$ and $\breve{\boldsymbol{P}}_+$ instead of $\lambda \boldsymbol{P}_+$.

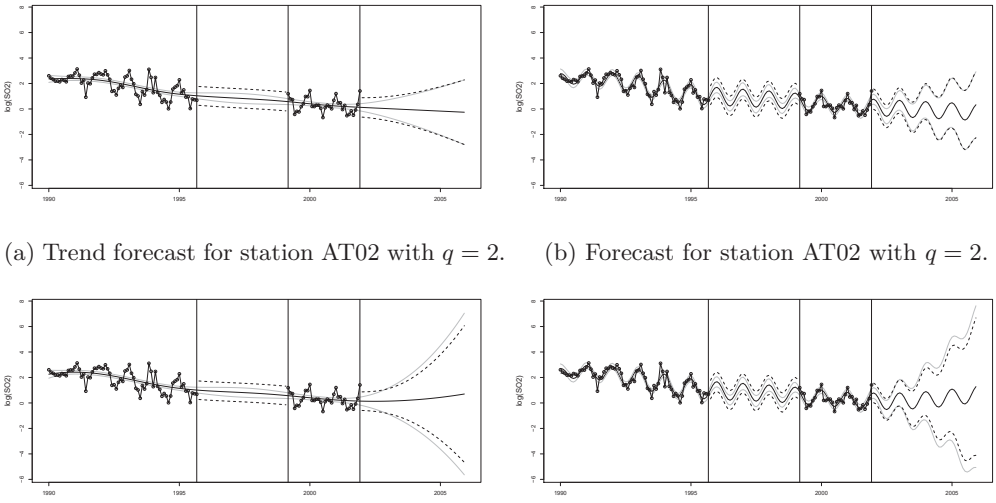
Keeping in mind that in a penalized spline modulation model, we have the trend and the seasonality components, it is straightforward to represent it as a mixed model.

In model (2.27) we consider a first order penalty for the modulation terms, since non-penalized terms for the modulation are $\cos(\omega \boldsymbol{x})$ and $\sin(\omega \boldsymbol{x})$. For instance, for $J = 1$ in (2.26) the fixed effect matrix for the smooth modulation model (2.27) is a design matrix of a harmonic regression model: $\breve{\boldsymbol{X}} = [\boldsymbol{1}_n \mid \boldsymbol{x} \mid \boldsymbol{x}^2 \mid ... \mid \boldsymbol{x}^{q-1} \mid \cos(\omega \boldsymbol{x}) \mid \sin(\omega \boldsymbol{x})]$.

The matrix of the random component, $\breve{\boldsymbol{Z}}$, is a block matrix: $\breve{\boldsymbol{Z}} = [\boldsymbol{Z} \mid \boldsymbol{CZ}_{cs} \mid \boldsymbol{SZ}_{cs}]$, where $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{\Omega}_r$ with $\boldsymbol{\Omega}_r = \boldsymbol{U}_r\tilde{\boldsymbol{\Sigma}}^{-1/2}$ ($\boldsymbol{U}_r$ and $\tilde{\boldsymbol{\Sigma}}$ are obtained from the SVD of $\boldsymbol{D}'_q\boldsymbol{D}_q$, with $\boldsymbol{D}_q$ a penalty matrix of order usually $q = 2$) and $\boldsymbol{Z}_{cs} = \boldsymbol{B}\breve{\boldsymbol{\Omega}}$, with $\breve{\boldsymbol{\Omega}} = \breve{\boldsymbol{U}}_r\breve{\tilde{\boldsymbol{\Sigma}}}^{-1/2}$ ($\breve{\boldsymbol{U}}_r$ and $\breve{\tilde{\boldsymbol{\Sigma}}}$ are obtained from the SVD of $\boldsymbol{D}'_{q_{cs}}\boldsymbol{D}_{q_{cs}}$, with $\boldsymbol{D}_{q_{cs}}$ a penalty matrix of order usually $q_{cs} = 1$).

The smoothing parameters λ, λ_c, λ_s in (2.28) become the ratio between the error variance term and the random effect variances σ^2, σ_c^2 and σ_s^2 for the trend and modulation terms respectively. The covariance matrix is a block diagonal matrix, $\breve{\boldsymbol{G}} = \text{blockdiag}(\boldsymbol{G}, \boldsymbol{G}_c, \boldsymbol{G}_s)$, where $\boldsymbol{G} = \sigma^2 \boldsymbol{I}_{c-q}$, $\boldsymbol{G}_c = \sigma_c^2 \boldsymbol{I}_{c-1}$ and $\boldsymbol{G}_s = \sigma_s^2 \boldsymbol{I}_{c-1}$.

The one-stage approach can deal with missing observations within and out-of-sample just by setting to zero the corresponding diagonal entries of $\boldsymbol{R}_+^{-1}$. Figure 5.1 shows the forecasted trend and final predictions (including the seasonal projections) for AT02 station with second and third penalty orders for the trend component. Notice that, as it is stated in Corollary 2.1, for second order penalty ($q = 2$) the trend forecast is linear, and for third order ($q = 3$) the trend forecast is quadratic.



(a) Trend forecast for station AT02 with $q = 2$.

(b) Forecast for station AT02 with $q = 2$.

(c) Trend forecast for station AT02 with $q = 3$.

(d) Forecast for station AT02 with $q = 3$.

Figure 2.6: Forecast for AT02 station. Top and bottom left figures show the data (points), the fitted and forecasted trend (black line), 95% confidence interval (grey lines) and 95% prediction interval (dashed lines) for second and third order penalties, respectively. Top and bottom right figures show the data (points), the fit and the forecast in the modulation model (black line), 95% confidence interval (grey lines) and 95% prediction interval (dashed lines) for second and third order penalties, respectively.

2.3.3 Forecasting sea level

As explained previously, the proposed methodology can also be used for errors with any variance-covariance matrix $\boldsymbol{R}$, for instance errors following an AR(1) with common variance and auto-correlation parameters σ_ϵ^2 and ρ, respectively. In order to illustrate this

we use a data set on sea level, global warning is a serious concern hence to know how fast sea levels are rising it is important to obtain predictions about the sea level. The dataset corresponds to the station of Vlissingen, a city in the southwestern Netherlands on the former island of Walcheren, the observed data consists of 157 observations, for year values between 1862 and 2018 (available at www.psmsl.org). We propose to consider the data between 1862 and 1998 as known and predict the sea level values for 20 out-of-sample years using the following two model:

$$y_i = f(x_i) + \epsilon_i, \tag{2.29}$$

where $\boldsymbol{x}$ are the years and $\boldsymbol{\epsilon}$ with mean zero and variance-covariance matrix the following matrix:

$$\boldsymbol{R} = \sigma^2_{\boldsymbol{\epsilon}} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}.$$

Building the extended basis and the penalty matrix analogously to the previous examples, it is straightforward to obtain the fit and the prediction applying (2.1). Figure 2.7 illustrates the fitted and forecasted trend and the 95% confidence interval (grey line) and 95% prediction interval (dashed line). As we can see, the values between 1998 and 2018 follow the predicted trend and lie inside the confidence interval.

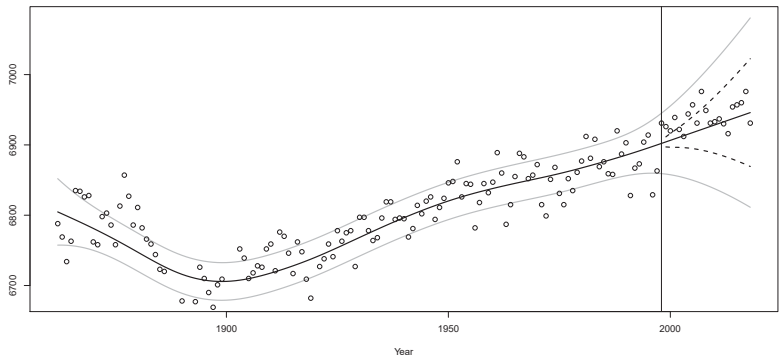


Figure 2.7: Fit, forecast, 95% confidence interval (grey lines) and 95% prediction interval (dashed lines) of a data set on the sea level. The vertical line indicates the year from which we predict, 1998.

In order to estimate the variance and correlation components we have modified the algorithm implemented in Rodríguez-Álvarez et al. (2018). The derivatives of the REML

and the R code can be seen in Appendix A.2. Notice that the idea behind the Schall (1991) algorithm is to speed up the estimation of the variance components since they deal with linear functions of the variance components. However, the REML derivatives are not linear with respect to the correlation parameter, so we need to solve nonlinear equations as it can be seen in Appendix A.2.

2.4 Memory of a P-spline

In some occasions, our knowledge of the data can influence our decision on the proportion of the data set that we want to use to predict new observations. Therefore, it may be important to know how much of the known information we are using to predict. In this section we introduce the concept of memory of a P-spline as a tool to provide that information and show some of its properties.

It is important to notice that, for i.i.d. errors, the matrix $\tilde{\boldsymbol{R}}_+^{-1}$ in (2.1) is a block diagonal matrix with entries zeros or ones. Therefore $\boldsymbol{H}_+$ in (2.1) has the following form:

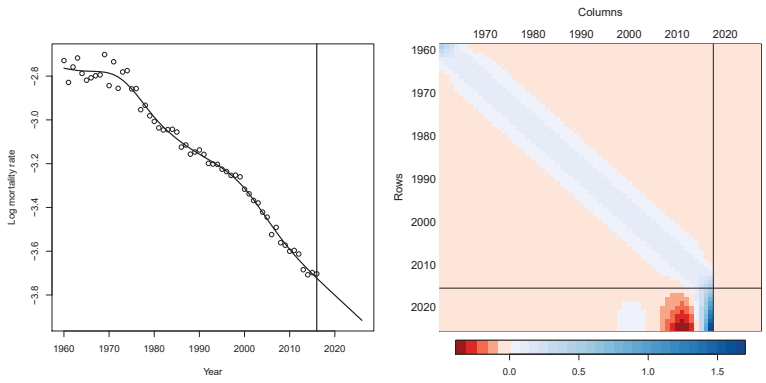
$$\boldsymbol{H}_+ = \begin{bmatrix} \boldsymbol{H} & \boldsymbol{O}_1 \\ \boldsymbol{H}_p & \boldsymbol{O}_2 \end{bmatrix}, \tag{2.30}$$

with $\boldsymbol{H}$ of size $n \times n$, $\boldsymbol{H}_p$ of size $n_p \times n$ and $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ matrices of zeros of size $n \times n_p$ and $n_p \times n_p$, respectively. I.e., the predicted values given by the missing value approach in the case of penalties based on differences between adjacent coefficients are

$$\hat{\boldsymbol{y}}_p = \boldsymbol{H}_p \boldsymbol{y},$$

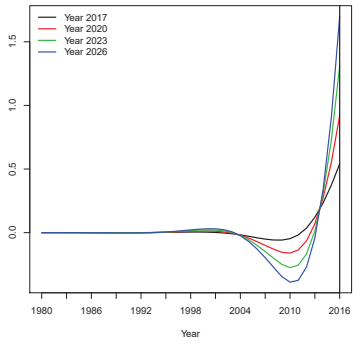
where $\boldsymbol{H}_p = \boldsymbol{B}_1(\boldsymbol{B}'\tilde{\boldsymbol{R}}^{-1}\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\tilde{\boldsymbol{R}}^{-1} - \boldsymbol{B}_2\boldsymbol{D}_2^{-1}\boldsymbol{D}_1(\boldsymbol{B}'\tilde{\boldsymbol{R}}^{-1}\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\tilde{\boldsymbol{R}}^{-1}$, with $\boldsymbol{B}$, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ as in (2.2) and $\boldsymbol{D}$, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ as in (2.4). Therefore, summarizing the rows and columns of $\boldsymbol{H}_p$ will give us an insight of how the past is affecting the prediction.

To illustrate the concept of *memory of a P-spline* we use the mortality data set of Section 2.1.1, a data set on log mortality rates of Spanish men aged 73 over the period 1960-2016. The data set contains 57 observations, i.e., the size of the hat matrix that give us the fit is 57×57. If we forecast up to 2026, i.e., we compute 10 new observations, the hat matrix $\boldsymbol{H}_p$ has size 10×57. Panel (a) of Figure 2.8 shows the fit and forecast of the log mortality rates until 2026. Panel (b) shows the image of the $\boldsymbol{H}_+$ matrix, and panel (c) dislplays the rows of $\boldsymbol{H}_p$. We can notice that all rows of $\boldsymbol{H}_p$ follow a similar pattern, i.e., if we consider each row as a function of year, we find that they



(a) Fit and forecast.

(b) $\boldsymbol{H}_+$ matrix.



(c) $\boldsymbol{H}_p$ matrix rows by column.

Figure 2.8: Panel (a): Fit and forecast. Panel (b): Image of $\boldsymbol{H}_+$. Panel (c): rows of $\boldsymbol{H}_p$.

behave similarly (see panel (c) of Figure 2.8). For instance, if the maximum of the last row is taken at the last column, this also happens in the rest of columns. Moreover, the contribution of each point in the past reduces gradually as we move away from the present. In particular, we could say that previous to 2001 the contribution is almost zero.

Based on the previous idea, we have developed the concept *memory of a P-spline*. This new idea will give us information on the overall weight of each observation on the prediction.

We have summarized the columns of $\boldsymbol{H}_p$ as follows: we add them (in absolute value) and standardize them by their sum, this will give us a vector of weights $\boldsymbol{w}$ of the same length of the response variable, $n \times 1$, we consider T as the number of steps backward from the last observation and associating the vector of weights to these values. Then, the ***memory of the P-spline*** is the 99^{th} percentile, t_0. That would mean that beyond t_0 steps backward no relevant information is affecting the prediction.

t	$\boldsymbol{w}_t$	t	$\boldsymbol{w}_t$	t	$\boldsymbol{w}_t$	t	$\boldsymbol{w}_t$
1	0.3293	7	0.0643	13	0.0061	19	0.0047
2	0.1894	8	0.0594	14	0.0013	20	0.0039
3	0.0851	9	0.0486	15	0.0031	21	0.0029
4	0.0162	10	0.0365	16	0.0031	22	0.0019
5	0.0364	11	0.0249	17	0.0046	23	0.0011
6	0.0577	12	0.0145	18	0.0050	24	0.0005

Table 2.1: Normalized weights, $\boldsymbol{w}_t$, for the number of steps backward from the last observed year.

Notice that defining the memory as the 99^{th} percentile is just one possible way to summarize the vector of weights. Summary statistics that treat the weights as if they are a discrete distribution (mean, quantiles, expectiles) are other choices.

To calculate the memory of the P-spline in the previous example, we compute the vector of weights, $\boldsymbol{w}$, shown in Table 2.1 (the values of $\boldsymbol{w}_t$ for $t = 25, ..., 57$ are not shown in the table since they are approximately 0), and obtain the 99^{th} percentile. In this case the memory of the P-spline is $t_0 = 18$, i.e., what has happened 18 years backward, before 1999, does not influence on the future.

Figure 2.9 illustrates the result. The left panel shows the vector of weights, the red line corresponds to the first year that influence the prediction, 1999. Right panel of Figure 2.9 shows the fit and the forecast of the log mortality rates until 2026, the data that are between the red and the black lines correspond to the data that contributes to the prediction.

2.4.1 Properties of the memory of a P-spline

In order to show the behaviour of the memory, we have performed a simulation study. We have applied the missing value approach with B-spline basis and second-order penalty to several simulated datasets by using different prediction horizons and bases of different sizes.

We have simulated from $y_i = f(x_i) + \epsilon_i$, $i = 1, ..., n$, $x_i \sim \text{Unif}[0, 5]$ with $n = 50$ and smooth functions and errors:

i) $f(x_i) = 2 + \sin\left(\frac{4\pi x_i}{5}\right)$, $\epsilon_i \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}} = 0.3)$.

ii) $f(x_i) = \exp(x_i)$, $\epsilon_i \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}} = 6)$.

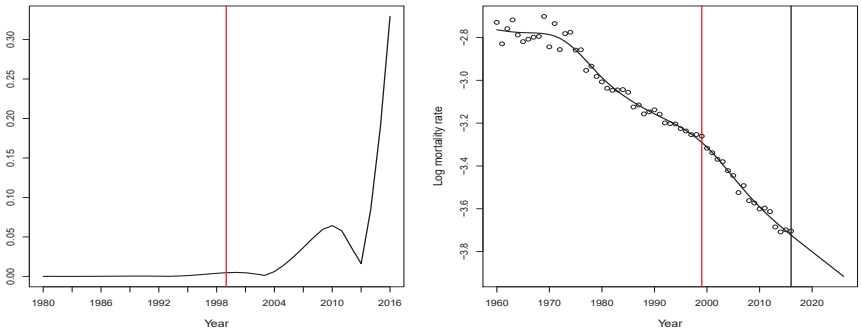


Figure 2.9: Left panel: vector of weights, the red line corresponds to the year from which we are using information, 1999. Right panel: fit and forecast of the log mortality rates until 2026, the data that are between the red and the black lines correspond to the data that contributes to the prediction.

iii) $f(x_i) = 2 + x_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}} = 0.6)$.

We have fitted each dataset using B-spline bases of different sizes, with dimensions $n \times c$, with $c = seq(40, 100, by = 10)$, and we have extended each one of the previous bases to predict until several horizons, we have used n_p values between 5 and 50 in steps of 5. Therefore, for each dataset, we have computed the memory for 7 sizes of the B-spline basis and 10 different prediction horizons, i.e. the memory for each function has been computed 70 times.

Figures 2.10, 2.11, 2.12 illustrate the results of the simulation study for three particular datasets, left panel shows the true function and the simulated data, and right panel shows the weight vectors for the different combinations of B-spline basis sizes and horizons. As we can see, the weight vectors for each dataset are almost equal, i.e. they do not depend on the B-spline basis sizes or on the prediction horizon. Moreover, it seems that the memory depends on the smoothness of the function from which we have generated the data. To corroborate this hypothesis, we have set several matrices $\boldsymbol{H}_+$, we have built them from a B-spline basis with size 50×25, and different smoothing parameters. For a sequence of λ values from 1 to 200 each 10 units, the memory is 8, 15, 18, 19, 21, 22, 23, 24, 25, 25, 26, 27, 27, 28, 28, 29, 29, 30, 30, 31, i.e. the memory increases as the smoothing parameter increases, as it was expected.

From the obtained results, we concluded:

1. The memory, like the effective dimension, does not depend on the size of the B-spline basis (provided that the basis is sufficiently large).

2. The memory does not depend on the prediction horizon.

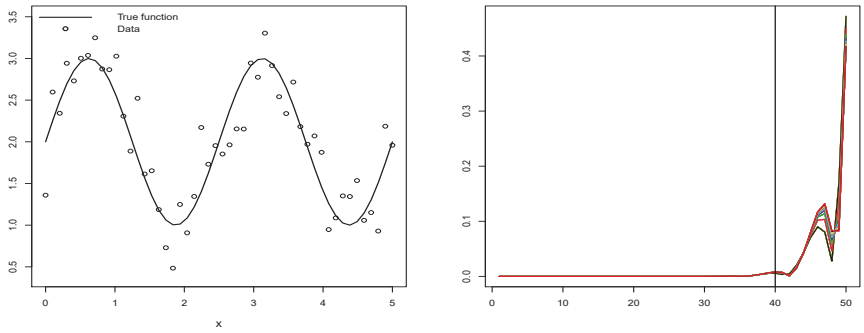


Figure 2.10: Left panel: simulated data from function i). Right panel: the associated 70 weight vectors for the different combinations of B-spline bases sizes and horizons, the vertical line indicates the memory of the P-spline, $t_0 = 11$.

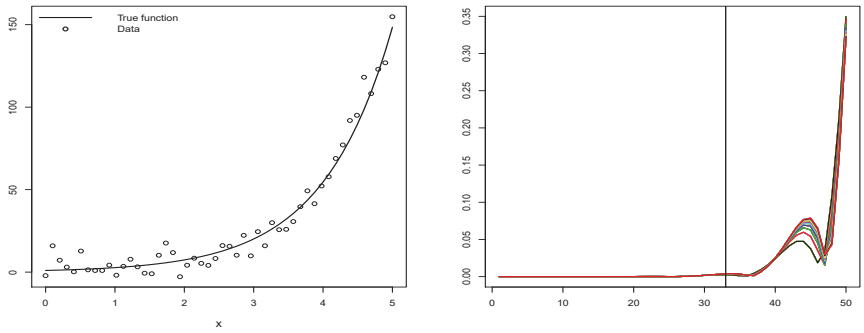


Figure 2.11: Left panel: simulated data from function ii). Right panel: the associated 70 weight vectors for the different combinations of B-spline bases sizes and horizons, the vertical line indicates the memory of the P-spline, $t_0 = 18$.

3. The memory depends on the smoothing parameter. The smaller (larger) the smoothing parameter is, the smaller (greater) the influence of the past on the predicted values is.

In order to illustrate property 2, we use the previous mortality data set. Figure 2.13 shows the vector of weights for different prediction horizons, as we can see the memory is always the same and data prior to 1999 do not contribute to the prediction. To illustrate the third property we fit and forecast up to 2026 the log mortality rates by using different smoothing parameters. Depending on the value of the smoothing parameter the memory is smaller or greater. In Figure 2.14, we can see that as the value of the smoothing parameter increases, the memory also increases.

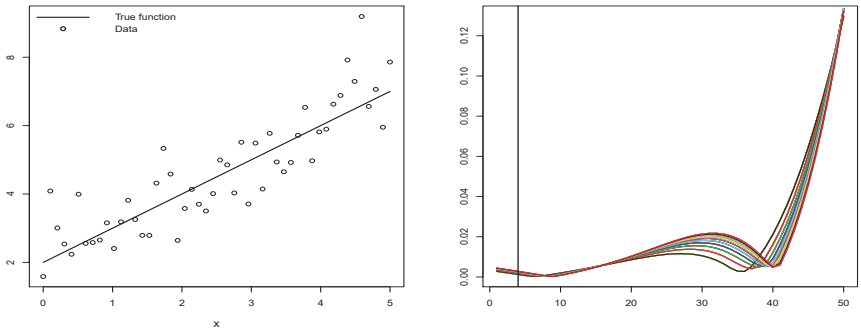


Figure 2.12: Left panel: simulated data from function iii). Right panel: the associated 70 weight vectors for the different combinations of B-spline bases sizes and horizons, the vertical line indicates the memory of the P-spline, $t_0 = 47$.

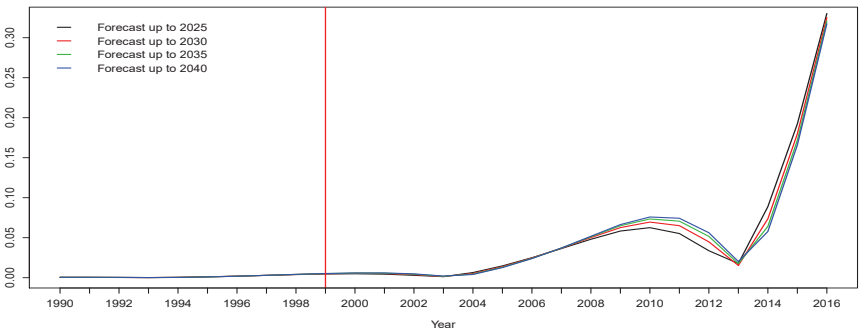


Figure 2.13: Vector of weights for different prediction horizons when we fit and forecast the log mortality rates of Spanish men aged 73.

2.5 Summary of the chapter

In this chapter, we have proposed a general framework for prediction of new observations in penalized regression, the methodology proposed can be accommodated to the different frameworks in which smoothing is carried out:

- Extend the basis used for regression and the penalty to control the smoothness in the framework of penalized regression based on quadratic penalties.
- Extend the fixed and random components in the context of mixed models.
- Define a Gaussian process for the extended set of random effects.

In the context of penalties based on differences between adjacent coefficients, we have proved the equivalence of all methods. We have also shown that the fit remains the

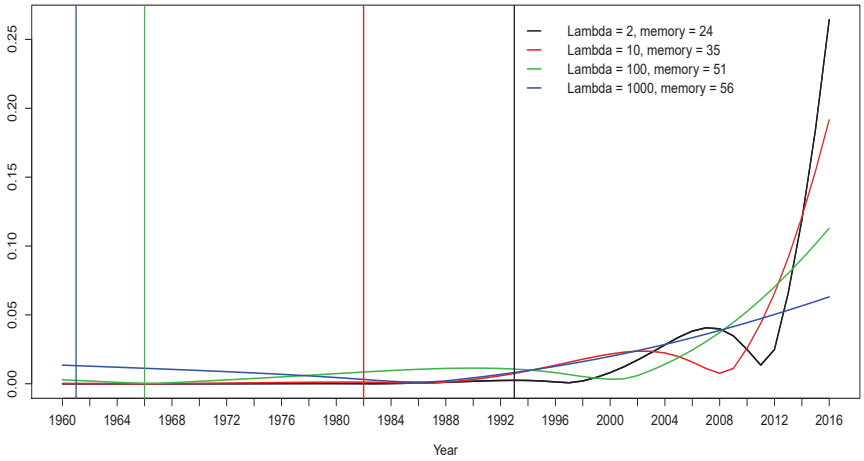


Figure 2.14: Vector of weights for different values of the smoothing parameter when we forecast 10 new observations of the log mortality rates of Spanish men aged 73.

same if the fit and the prediction are obtained simultaneously and that the order of the penalty function, which is less relevant in the smoothing of data, is now critical, because the penalty function determines the shape of the prediction. This is due to the fact that, for a given penalty order q, the coefficients that determine the prediction are a linear combination of order $q-1$ of the last q coefficients that determine the fit. With regard to the smoothing parameter, we have proved that the solution of the REML and of the extended REML are the same, i.e. the smoothing parameter estimated in the fit is the same as the smoothing parameter used to fit and predict simultaneously.

We have also introduced the concept of "memory of a P-spline" as a tool to know how much known information from the past we are using to predict. Through a simulation study we have been able to conclude that the memory just depends on the smoothing parameter, provided that the regression basis is sufficiently large. To illustrate the methodology, the proved results and the concept of "memory of a P-spline", we have showed the performance of the method proposed with B-spline basis and penalties based on differences by using four examples based on real data sets. A data set on the log mortality rates of Spanish men allow us to illustrate the proved properties and the concept of "memory of a P-spline", with the aim of predicting the aboveground biomass of trees, we show an example of predicting within the framework of additive models and in the case when out-of-sample prediction is needed to the left and right of the interval where the covariate is observed. We also show classic examples where forecasting is needed with datasets collected over time, with real datasets where the errors can be correlated.

Chapter 3

Out-of-sample prediction in multidimensional P-spline models

In this chapter, we propose a general framework for out-of-sample prediction in multidimensional smoothing. The need for out-of-sample prediction in two or more dimensions appears in contexts such as spatial or spatio-temporal modelling, where we can be interested in prediction at new locations. This would involve out-of-sample prediction for two or three covariates (latitude, longitude and time). To achieve this goal we extend the proposal of Currie et al. (2004) to predict when the two covariates are extended.

As we have seen in the previous chapter, in one dimension, doing the out-of-sample prediction in one or two steps has no influence on the fit to the data, but this is not true when models include interaction terms. Studying the properties of the method proposed by Currie et al. (2004), we will see that the fit changes when the fit and the prediction are obtained simultaneously. We will show that, for the particular case in which just one of the two covariates is extended, the fit can be maintained by modifying the extended penalty matrix. However, when the two covariates are extended the penalty matrix can not be modified, since the matrices involved in obtaining the estimated parameters become singular. As a general solution to ensure the invariance of the fit we will impose restrictions on the coefficients. We will achieve it by using the Lagrange formulation of the least squares minimization problem following Greene and Seaks (1991).

This chapter is organized as follows. In Section 3.1 we briefly review the P-splines methodology for the multidimensional case and its reparameterization as mixed models. Section 3.2.1 is dedicated to extend the proposal of Currie et al. (2004) to the case in which out-of-sample prediction is needed in both covariates of the interaction term. Then, we show the properties satisfied, under certain conditions, by the coefficients that determine the prediction. Furthermore, in Section 3.2.2 we propose a method, based on Lagrange multipliers, to obtain constrained predictions. Section 3.2.3 provides several

examples in order to illustrate how restrictions can also solve different situations in which constrained prediction is needed. In particular, we will show how to solve the crossover problem of adjacent ages when mortality tables are forecasted. Finally, Section 3.3 shows how out-of-sample predictions can be carried out in the context of multidimensional smooth mixed models. We propose different reparametrizations to predict new values and also show how to impose constraints in this context.

3.1 P-splines and mixed models representation for multidimensional data

The easiest approach to handle more than one covariates in the context of smoothing models, is to use additive models (they do not include interaction between terms). In that case, everything related to prediction under the Gaussian framework has been done in the previous chapter. Therefore, here we focus on the case of interactions, i.e., in additive models that include terms of the form $f(\boldsymbol{z}, \boldsymbol{x})$, where $\boldsymbol{z}$ and $\boldsymbol{x}$ are the covariates.

In order to study the prediction approach given in Currie et al. (2004), we briefly review the P-splines methodology and its reparameterization as a mixed model in the two-dimensional case.

3.1.1 Multidimensional P-splines

We consider a general non-parametric two-dimensional regression model:

$$\boldsymbol{y} = f(\boldsymbol{z}, \boldsymbol{x}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \backsim N(\boldsymbol{0}, \boldsymbol{R}), \tag{3.1}$$

where $\boldsymbol{z}$, $\boldsymbol{x}$ are the regressors, $\boldsymbol{R} = \sigma^2_{\boldsymbol{\epsilon}} \boldsymbol{I}$, i.e., $\boldsymbol{\epsilon}$ are independent and identically distributed errors with variance $\sigma^2_{\boldsymbol{\epsilon}}$, and $f(\cdot)$ is a $2-$multidimensional smooth function that depends on the 2 explanatory variables $\boldsymbol{z} = (z_1, ..., z_{n_{\boldsymbol{z}}})'$ and $\boldsymbol{x} = (x_1, ..., x_{n_{\boldsymbol{x}}})'$, each of them of lengths $n_{\boldsymbol{z}}$ and $n_{\boldsymbol{x}}$, respectively. Notice that for scattered data $n_{\boldsymbol{z}} = n_{\boldsymbol{x}}$ while for array data $n_{\boldsymbol{z}}$ and $n_{\boldsymbol{x}}$ can have different values. Although we are assuming i.i.d. errors for simplicity the results can be easily extended to the case of a general variance-covariance matrix $\boldsymbol{R}$, as it was shown in the previous chapter. Suppose now that we are interested in fitting model (3.1), and assume that the function $f(\boldsymbol{z}, \boldsymbol{x})$ can be represented in terms of basis functions:

$$f(\boldsymbol{z}, \boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\theta}, \tag{3.2}$$

with $\boldsymbol{B}$ a B-spline regression basis, and $\boldsymbol{\theta}$ the vector of coefficients. If we consider array data, the smooth multidimensional surface is constructed from the Kronecker product of

the marginal B-spline basis for each covariate, the basis for the model (3.2) is

$$\boldsymbol{B} = \boldsymbol{B_x} \otimes \boldsymbol{B_z}, \tag{3.3}$$

where $\otimes$ is the Kronecker product of two matrices, and $\boldsymbol{B_x} = \boldsymbol{B}(\boldsymbol{x})$ and $\boldsymbol{B_z} = \boldsymbol{B}(\boldsymbol{z})$, of dimensions $n_{\boldsymbol{x}} \times c_{\boldsymbol{x}}$ and $n_{\boldsymbol{z}} \times c_{\boldsymbol{z}}$, are the marginal B-spline basis for $\boldsymbol{x}$ and $\boldsymbol{z}$, respectively. Then, the dimension of (3.3) is $n_{\boldsymbol{x}} n_{\boldsymbol{z}} \times c_{\boldsymbol{x}} c_{\boldsymbol{z}}$. On the other hand, if we consider scattered data, the basis is constructed from the Tensor product of marginal B-spline basis defined in Currie et al. (2006) as the Box-Product, denoted by symbol $\square$:

$$\boldsymbol{B} = \boldsymbol{B_x} \square \boldsymbol{B_z} = (\boldsymbol{B_x} \otimes \mathbf{1}'_{c_{\boldsymbol{z}}}) \odot (\mathbf{1}'_{c_{\boldsymbol{x}}} \otimes \boldsymbol{B_z}),$$

where the operator $\odot$ is the element-wise matrix product and $\mathbf{1}_{c_{\boldsymbol{z}}}$ and $\mathbf{1}_{c_{\boldsymbol{x}}}$ are column vectors of ones of lengths $c_{\boldsymbol{z}}$ and $c_{\boldsymbol{x}}$.

In both cases, the vector of coefficients $\boldsymbol{\theta}$ can be arranged into a $c_{\boldsymbol{z}} \times c_{\boldsymbol{x}}$ matrix $\boldsymbol{\Theta}$, that is

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\theta}_{11} & \boldsymbol{\theta}_{12} & \cdots & \boldsymbol{\theta}_{1c_{\boldsymbol{x}}} \\ \boldsymbol{\theta}_{21} & \boldsymbol{\theta}_{22} & \cdots & \boldsymbol{\theta}_{2c_{\boldsymbol{x}}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\theta}_{c_{\boldsymbol{z}}1} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}2} & \cdots & \boldsymbol{\theta}_{c_{\boldsymbol{z}}c_{\boldsymbol{x}}} \end{bmatrix}, \tag{3.4}$$

then, the two-dimensional P-spline model can be written as

$$f(\boldsymbol{z}, \boldsymbol{x}) = (\boldsymbol{B_x} \otimes \boldsymbol{B_z})\,\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{B_z} \boldsymbol{\Theta} \boldsymbol{B}'_{\boldsymbol{x}}),$$

where $\mathrm{vec}(\cdot)$ denotes the vectorization operator.

In the two dimensional case, the penalty on the coefficients vector $\boldsymbol{\theta}$ penalizes the difference between adjacent coefficients of rows and columns of the matrix $\boldsymbol{\Theta}$, defined in (3.4). The penalty on rows of $\boldsymbol{\Theta}$ is:

$$\sum_{j=1}^{c_{\boldsymbol{z}}} \boldsymbol{\theta}'_j \boldsymbol{D}'_{\boldsymbol{z}} \boldsymbol{D}_{\boldsymbol{z}} \boldsymbol{\theta}_j = \boldsymbol{\theta}' (\boldsymbol{I}_{c_{\boldsymbol{x}}} \otimes \boldsymbol{D}'_{\boldsymbol{z}} \boldsymbol{D}_{\boldsymbol{z}}) \boldsymbol{\theta},$$

and, similarly, on the columns:

$$\sum_{i=1}^{c_{\boldsymbol{x}}} \boldsymbol{\theta}'_i \boldsymbol{D}'_{\boldsymbol{x}} \boldsymbol{D}_{\boldsymbol{x}} \boldsymbol{\theta}_i = \boldsymbol{\theta}' (\boldsymbol{D}'_{\boldsymbol{x}} \boldsymbol{D}_{\boldsymbol{x}} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}}) \boldsymbol{\theta},$$

where $\boldsymbol{D_z}$ and $\boldsymbol{D_x}$ are the difference matrices acting on the rows and columns of $\boldsymbol{\Theta}$,

respectively. Therefore, the penalty matrix $\boldsymbol{P}$ in two dimensions is:

$$\boldsymbol{P} = \lambda_{\boldsymbol{z}} \underbrace{\boldsymbol{I}_{c_{\boldsymbol{x}}} \otimes \boldsymbol{D}_{\boldsymbol{z}}' \boldsymbol{D}_{\boldsymbol{z}}}_{\boldsymbol{P}^{\boldsymbol{z}}} + \lambda_{\boldsymbol{x}} \underbrace{\boldsymbol{D}_{\boldsymbol{x}}' \boldsymbol{D}_{\boldsymbol{x}} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}}}_{\boldsymbol{P}^{\boldsymbol{x}}}, \tag{3.5}$$

where $\lambda_{\boldsymbol{z}}$ and $\lambda_{\boldsymbol{x}}$ are the smoothing parameters for each dimension of the model. Since $\lambda_{\boldsymbol{z}}$ and $\lambda_{\boldsymbol{x}}$ are not necessary equal, the penalty (3.5) allows for anisotropic smoothing. To estimate the coefficients, Eilers and Marx (1996) minimize the penalized sum of squares:

$$S(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\theta})'(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\theta}) + \boldsymbol{\theta}' \boldsymbol{P} \boldsymbol{\theta}. \tag{3.6}$$

Therefore, for given values of $\lambda_{\boldsymbol{z}}$ and $\lambda_{\boldsymbol{x}}$, the solution of the penalized sum of squares (3.6), is:

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{P})^{-1} \boldsymbol{B}' \boldsymbol{y}. \tag{3.7}$$

The smoothing parameter of each dimension can be estimated using an information criteria (such as Akaike or Bayesian criteria) or a cross-validation criteria method.

Once we have presented a brief introduction of the multidimensional P-splines, in the next section, we detail its representation as mixed model. Although we will use B-spline basis and penalties based on differences, the methodology proposed here can be extended to any basis and quadratic penalty.

3.1.2 Multidimensional representation of P-splines as mixed models

In order to extend the prediction methodology to the two-dimensional case, we use the two dimensional mixed formulation of P-splines. Here we give a short summary, for more detailed description, see Lee (2010).

As in the univariate case, we have to define a transformation matrix $\boldsymbol{\Omega}$ that allow us to rewrite the regression basis $\boldsymbol{B} = \boldsymbol{B}_{\boldsymbol{x}} \otimes \boldsymbol{B}_{\boldsymbol{z}}$ as the mixed model matrices $[\boldsymbol{X} \mid \boldsymbol{Z}]$ and its associated regression coefficients $\boldsymbol{\theta}$ as the mixed model coefficients $[\beta' \mid \boldsymbol{\alpha}']'$. For that, we follow Lee (2010) and consider the SVD of the marginal matrices $\boldsymbol{D}_{\boldsymbol{x}}' \boldsymbol{D}_{\boldsymbol{x}}$ and $\boldsymbol{D}_{\boldsymbol{z}}' \boldsymbol{D}_{\boldsymbol{z}}$:

$$\boldsymbol{D}_i' \boldsymbol{D}_i = \boldsymbol{U}_i \boldsymbol{\Sigma}_i \boldsymbol{U}_i',$$

with $\boldsymbol{U}_i$ the matrix of eigenvectors that, as in the univariate case, can be splitted into two sub-matrices, $\boldsymbol{U}_i = [\boldsymbol{U}_{if} \mid \boldsymbol{U}_{ir}]$ ($\boldsymbol{U}_{if}$ spanning the null space and $\boldsymbol{U}_{ir}$ spanning the non-null space), and $\boldsymbol{\Sigma}_i$ a diagonal matrix containing the eigenvalues, for $i = \boldsymbol{z}, \boldsymbol{x}$.

Therefore, we consider the following transformation matrix $\boldsymbol{\Omega}$:

$$\boldsymbol{\Omega} = [\underbrace{\boldsymbol{U}_{\boldsymbol{x}f} \otimes \boldsymbol{U}_{\boldsymbol{z}f}}_{\boldsymbol{\Omega}_f} \mid \underbrace{\boldsymbol{U}_{\boldsymbol{x}r} \otimes \boldsymbol{U}_{\boldsymbol{z}f} \mid \boldsymbol{U}_{\boldsymbol{x}f} \otimes \boldsymbol{U}_{\boldsymbol{z}r} \mid \boldsymbol{U}_{\boldsymbol{x}r} \otimes \boldsymbol{U}_{\boldsymbol{z}r}}_{\boldsymbol{\Omega}_r}], \tag{3.8}$$

which is obtained by rearranging the block matrices as $[\boldsymbol{U}_{\boldsymbol{x}f} \mid \boldsymbol{U}_{\boldsymbol{x}r}] \otimes [\boldsymbol{U}_{\boldsymbol{z}f} \mid \boldsymbol{U}_{\boldsymbol{z}r}]$. Given $\boldsymbol{\Omega}$ in (3.8), the mixed model matrices are:

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{\Omega}_f = (\boldsymbol{B}_{\boldsymbol{x}} \otimes \boldsymbol{B}_{\boldsymbol{z}})(\boldsymbol{U}_{\boldsymbol{x}f} \otimes \boldsymbol{U}_{\boldsymbol{z}f}) = \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}f} \otimes \boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}f},$$

and

$$\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{\Omega}_r = [\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}r} \otimes \boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}f} \mid \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}f} \otimes \boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}r} \mid \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}r} \otimes \boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}r}]\,.$$

Defining $\boldsymbol{X}_i = \boldsymbol{B}_i\boldsymbol{U}_{if}$ and $\boldsymbol{Z}_i = \boldsymbol{B}_i\boldsymbol{U}_{ir}$ $(i = \boldsymbol{z}, \boldsymbol{x})$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ can be written as:

$$\begin{aligned}\boldsymbol{X} &= \boldsymbol{X}_{\boldsymbol{x}} \otimes \boldsymbol{X}_{\boldsymbol{z}},\\ \mathbf{Z} &= [\mathbf{Z}_{\boldsymbol{x}} \otimes \boldsymbol{X}_{\boldsymbol{z}} \mid \boldsymbol{X}_{\boldsymbol{x}} \otimes \boldsymbol{Z}_{\boldsymbol{z}} \mid \boldsymbol{Z}_{\boldsymbol{x}} \otimes \boldsymbol{Z}_{\boldsymbol{z}}].\end{aligned}$$

Notice that the capital letters $\boldsymbol{X}$ and $\boldsymbol{Z}$ denote the model matrices associated to the fixed and random effects, and the subscript letters, $\boldsymbol{x}$ and $\boldsymbol{z}$, the covariate to which the matrices are associated. Therefore, the mixed model coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are $\boldsymbol{\theta}$ as $\boldsymbol{\beta} = \boldsymbol{\Omega}_f'\boldsymbol{\theta}$ and $\boldsymbol{\alpha} = \boldsymbol{\Omega}_r'\boldsymbol{\theta}$. Moreover, for the penalty matrix given in (3.5) and the transformation matrix given in (3.8), the mixed model precision matrix is:

$$\boldsymbol{F} = \begin{bmatrix} \lambda_{\boldsymbol{x}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \otimes \boldsymbol{I}_{q_{\boldsymbol{z}}} & & \\ & \lambda_{\boldsymbol{z}}\boldsymbol{I}_{q_{\boldsymbol{x}}} \otimes \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} & \\ & & \lambda_{\boldsymbol{x}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}-q_{\boldsymbol{z}}} + \lambda_{\boldsymbol{x}}\boldsymbol{I}_{c_{\boldsymbol{x}}-q_{\boldsymbol{x}}} \otimes \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} \end{bmatrix},$$

where the matrices $\tilde{\boldsymbol{\Sigma}}_i$ $(i = \boldsymbol{z}, \boldsymbol{x})$ were defined previously. Therefore, the variance-covariance matrix $\boldsymbol{G}$ is:

$$\boldsymbol{G} = \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{F}^{-1}.$$

Using the previous mixed model matrices and random effects covariance matrix, the estimation procedure can be carried out as it was shown in Section 1.2.2.

3.2 Prediction in additive models based on multidimensional penalized splines

In this section, we extend the approach given in Currie et al. (2004) to obtain the prediction when not only one of the two independent variables but the two extend. For the particular case in which a single covariate is extended, we state several properties of the prediction method. Since natural extensions of penalty matrices provide changes in the fit, to overcome this problem we propose the use of restrictions.

3.2.1 Out-of-sample prediction

In the framework of model (3.1), considering array data and a vector of $n_{\boldsymbol{z}} n_{\boldsymbol{x}} \times 1$ observations, $\boldsymbol{y}$, of the response variable, suppose that we want to predict $n_p = n_{\boldsymbol{z}} n_{\boldsymbol{x}_p} + n_{\boldsymbol{z}_p} n_{\boldsymbol{x}} + n_{\boldsymbol{z}_p} n_{\boldsymbol{x}_p}$ new values at $(\boldsymbol{z}, \boldsymbol{x}_p)$, $(\boldsymbol{z}_p, \boldsymbol{x})$ and $(\boldsymbol{z}_p, \boldsymbol{x}_p)$, i.e., if we arrange the observations vector into a matrix $\boldsymbol{Y}$ of dimension $n_{\boldsymbol{z}} \times n_{\boldsymbol{x}}$, the observed and predicted values can be arranged into a matrix of dimension $n_{\boldsymbol{z}_+} \times n_{\boldsymbol{x}_+}$ ($n_{\boldsymbol{z}_+} = n_{\boldsymbol{z}} + n_{\boldsymbol{z}_p}$, $n_{\boldsymbol{x}_+} = n_{\boldsymbol{x}} + n_{\boldsymbol{x}_p}$), as:

$$\boldsymbol{Y}_+ = \begin{bmatrix} \boldsymbol{Y} & \boldsymbol{Y}_{\boldsymbol{z}\boldsymbol{x}_p} \\ \boldsymbol{Y}_{\boldsymbol{z}_p\boldsymbol{x}} & \boldsymbol{Y}_{\boldsymbol{z}_p\boldsymbol{x}_p} \end{bmatrix}. \tag{3.9}$$

Notice that the dimensions of $\boldsymbol{Y}_{\boldsymbol{z}\boldsymbol{x}_p}$, $\boldsymbol{Y}_{\boldsymbol{z}_p\boldsymbol{x}}$ and $\boldsymbol{Y}_{\boldsymbol{z}_p\boldsymbol{x}_p}$ are $n_{\boldsymbol{z}} \times n_{\boldsymbol{x}_p}$, $n_{\boldsymbol{z}_p} \times n_{\boldsymbol{x}}$ and $n_{\boldsymbol{z}_p} \times n_{\boldsymbol{x}_p}$, respectively.

We propose to fit and predict the model simultaneously considering the following extended model:

$$\boldsymbol{y}_+ = \boldsymbol{B}_+\boldsymbol{\theta}_+ + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+) \tag{3.10}$$

where $\boldsymbol{y}_+ = \text{vec}(\boldsymbol{Y}_+)$, with $\boldsymbol{Y}_+$ as in (3.9), where $\boldsymbol{Y}$ are the observed values and $\boldsymbol{Y}_{\boldsymbol{z}\boldsymbol{x}_p}$, $\boldsymbol{Y}_{\boldsymbol{z}_p\boldsymbol{x}}$ and $\boldsymbol{Y}_{\boldsymbol{z}_p\boldsymbol{x}_p}$ are arbitrary values, and $\boldsymbol{R}_+ = \sigma^2_{\boldsymbol{\epsilon}} \tilde{\boldsymbol{R}}_+$ with $\tilde{\boldsymbol{R}}_+ = \tilde{\boldsymbol{R}}_{\boldsymbol{x}_+} \otimes \tilde{\boldsymbol{R}}_{\boldsymbol{z}_+}$, and $\tilde{\boldsymbol{R}}_{\boldsymbol{x}_+}$ and $\tilde{\boldsymbol{R}}_{\boldsymbol{z}_+}$ diagonal matrices of dimensions $n_{\boldsymbol{x}_+} \times n_{\boldsymbol{x}_+}$ ($n_{\boldsymbol{x}_+} = n_{\boldsymbol{x}} + n_{\boldsymbol{x}_p}$) and $n_{\boldsymbol{z}_+} \times n_{\boldsymbol{z}_+}$ ($n_{\boldsymbol{z}_+} = n_{\boldsymbol{z}} + n_{\boldsymbol{z}_p}$), respectively, with infinity entries if the data is missing and 1 if the data is observed. The quantity infinity expresses that we do not have any information about the data $\boldsymbol{y}_p$. The extended basis is the Kronecker product of the two extended marginal B-spline basis, $\boldsymbol{B}_{\boldsymbol{x}_+} = \boldsymbol{B}(\boldsymbol{x}_+)$ and $\boldsymbol{B}_{\boldsymbol{z}_+} = \boldsymbol{B}(\boldsymbol{z}_+)$, of dimensions $n_{\boldsymbol{x}_+} \times c_{\boldsymbol{x}_+}$ and $n_{\boldsymbol{z}_+} \times c_{\boldsymbol{z}_+}$, respectively:

$$\boldsymbol{B}_+ = \boldsymbol{B}_{\boldsymbol{x}_+} \otimes \boldsymbol{B}_{\boldsymbol{z}_+} = \begin{bmatrix} \boldsymbol{B}_{\boldsymbol{x}} & \boldsymbol{O} \\ \boldsymbol{B}_{\boldsymbol{x}_{(1)}} & \boldsymbol{B}_{\boldsymbol{x}_{(2)}} \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{B}_{\boldsymbol{z}} & \boldsymbol{O} \\ \boldsymbol{B}_{\boldsymbol{z}_{(1)}} & \boldsymbol{B}_{\boldsymbol{z}_{(2)}} \end{bmatrix},$$

where the extended bases $\boldsymbol{B}_{\boldsymbol{x}_+}$ and $\boldsymbol{B}_{\boldsymbol{z}_+}$ are built from a new set of knots that consists

of the original knots and extended to cover the full range of $\boldsymbol{x}_+$ and $\boldsymbol{z}_+$, respectively.

To estimate the extended coefficients, we minimize the following function of $\boldsymbol{\theta}_+$:

$$S(\boldsymbol{\theta}_+) = (\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_+)'\tilde{\boldsymbol{R}}_+^{-1}(\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_+) + \boldsymbol{\theta}'_+\boldsymbol{P}_+\boldsymbol{\theta}_+, \tag{3.11}$$

with extended penalty matrix

$$\boldsymbol{P}_+ = \lambda_{\boldsymbol{z}}\boldsymbol{P}_+^{\boldsymbol{z}_+} + \lambda_{\boldsymbol{x}}\boldsymbol{P}_+^{\boldsymbol{x}_+}, \tag{3.12}$$

where $\lambda_{\boldsymbol{z}}$ and $\lambda_{\boldsymbol{x}}$, and $\boldsymbol{P}_+^{\boldsymbol{z}_+}$ and $\boldsymbol{P}_+^{\boldsymbol{x}_+}$ are the smoothing parameters and the extended penalty matrices for each dimension of the model, respectively. For the particular case of penalties based on differences, we consider:

$$\begin{aligned}\boldsymbol{P}_+^{\boldsymbol{z}_+} &= \boldsymbol{I}_{c_{\boldsymbol{x}_+}} \otimes \boldsymbol{D}'_{\boldsymbol{z}_+}\boldsymbol{D}_{\boldsymbol{z}_+} = \begin{bmatrix}\boldsymbol{I}_{c_{\boldsymbol{x}}} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{I}_{c_{\boldsymbol{x}_p}}\end{bmatrix} \otimes \boldsymbol{D}'_{\boldsymbol{z}_+}\boldsymbol{D}_{\boldsymbol{z}_+} = \begin{bmatrix}\boldsymbol{I}_{c_{\boldsymbol{x}}} \otimes \boldsymbol{D}'_{\boldsymbol{z}_+}\boldsymbol{D}_{\boldsymbol{z}_+} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{I}_{c_{\boldsymbol{x}_p}} \otimes \boldsymbol{D}'_{\boldsymbol{z}_+}\boldsymbol{D}_{\boldsymbol{z}_+}\end{bmatrix}\\ &= \begin{bmatrix}\boldsymbol{P}_{+11}^{\boldsymbol{z}_+} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{P}_{+22}^{\boldsymbol{z}_+}\end{bmatrix},\end{aligned} \tag{3.13}$$

and

$$\begin{aligned}\boldsymbol{P}_+^{\boldsymbol{x}_+} &= \boldsymbol{D}_{\boldsymbol{x}_+}'\boldsymbol{D}_{\boldsymbol{x}_+} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}} = \begin{bmatrix}(\boldsymbol{D}'_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{x}} + \boldsymbol{D}'_{\boldsymbol{x}(1)}\boldsymbol{D}_{\boldsymbol{x}(1)}) \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}} & \boldsymbol{D}'_{\boldsymbol{x}(1)}\boldsymbol{D}_{\boldsymbol{x}(2)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}}\\ \boldsymbol{D}'_{\boldsymbol{x}(2)}\boldsymbol{D}_{\boldsymbol{x}(1)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}} & \boldsymbol{D}'_{\boldsymbol{x}(2)}\boldsymbol{D}_{\boldsymbol{x}(2)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}}\end{bmatrix}\\ &= \begin{bmatrix}\boldsymbol{P}_{+11}^{\boldsymbol{x}_+} & \boldsymbol{P}_{+12}^{\boldsymbol{x}_+}\\ \boldsymbol{P}_{+21}^{\boldsymbol{x}_+} & \boldsymbol{P}_{+22}^{\boldsymbol{x}_+}\end{bmatrix},\end{aligned} \tag{3.14}$$

where $\boldsymbol{D}_{\boldsymbol{x}_+}$ and $\boldsymbol{D}_{\boldsymbol{z}_+}$ are the difference matrices acting on the columns and rows of the matrix formed by the extended vector of coefficients $\boldsymbol{\theta}_+$. Notice that $\boldsymbol{D}_{\boldsymbol{z}_+}$ and $\boldsymbol{D}_{\boldsymbol{x}_+}$ are direct extensions of $\boldsymbol{D}_{\boldsymbol{z}}$ and $\boldsymbol{D}_{\boldsymbol{x}}$ but $\boldsymbol{P}_+^{\boldsymbol{z}_+}$ and $\boldsymbol{P}_+^{\boldsymbol{x}_+}$ are not direct extensions of $\boldsymbol{P}^{\boldsymbol{z}}$ and $\boldsymbol{P}^{\boldsymbol{x}}$. Moreover, if $\boldsymbol{\theta}$ are the coefficients that determine the fit arranged in a matrix of dimension $c_{\boldsymbol{z}} \times c_{\boldsymbol{x}}$ and we are extending the two covariates, $\boldsymbol{\theta}_+ = \text{vec}(\boldsymbol{\Theta}_+)$, with $\boldsymbol{\Theta}_+$:

$$\boldsymbol{\Theta}_+ = \left[\begin{array}{cccccccc}
\boldsymbol{\theta}_{11} & \boldsymbol{\theta}_{12} & \cdots & \boldsymbol{\theta}_{1\ c_{\boldsymbol{x}}-1} & \boldsymbol{\theta}_{1c_{\boldsymbol{x}}} & \boldsymbol{\theta}_{1\ c_{\boldsymbol{x}}+1} & \boldsymbol{\theta}_{1\ c_{\boldsymbol{x}}+2} & \cdots\\
\boldsymbol{\theta}_{21} & \boldsymbol{\theta}_{22} & \cdots & \boldsymbol{\theta}_{2\ c_{\boldsymbol{x}}-1} & \boldsymbol{\theta}_{2c_{\boldsymbol{x}}} & \boldsymbol{\theta}_{2\ c_{\boldsymbol{x}}+1} & \boldsymbol{\theta}_{2\ c_{\boldsymbol{x}}+2} & \cdots\\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots\\
\boldsymbol{\theta}_{c_{\boldsymbol{z}}1} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}2} & \cdots & \boldsymbol{\theta}_{c_{\boldsymbol{z}}\ c_{\boldsymbol{x}}-1} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}c_{\boldsymbol{x}}} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}\ c_{\boldsymbol{x}}+1} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}\ c_{\boldsymbol{x}}+2} & \cdots\\
\hline
\boldsymbol{\theta}_{c_{\boldsymbol{z}}+1\ 1} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}+1\ 2} & \cdots & \boldsymbol{\theta}_{c_{\boldsymbol{z}}+1\ c_{\boldsymbol{x}}-1} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}+1\ c_{\boldsymbol{x}}} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}+1\ c_{\boldsymbol{x}}+1} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}+1\ c_{\boldsymbol{x}}+2} & \cdots\\
\boldsymbol{\theta}_{c_{\boldsymbol{z}}+2\ 1} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}+2\ 2} & \cdots & \boldsymbol{\theta}_{c_{\boldsymbol{z}}+2\ c_{\boldsymbol{x}}-1} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}+2\ c_{\boldsymbol{x}}} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}+2\ c_{\boldsymbol{x}}+1} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}+2\ c_{\boldsymbol{x}}+2} & \cdots\\
\vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \cdots
\end{array}\right],$$

where the coefficients in black are the ones that determine the fit and the coefficients in blue the ones that determine the prediction.

The solution of the extended penalized least squares problem (3.11) is:

$$\boldsymbol{\theta}_+ = (\boldsymbol{B}'_+\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{B}_+ + \boldsymbol{P}_+)^{-1}\boldsymbol{B}'_+\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{y}_+.$$

As mentioned earlier, an information criteria or a cross-validation criteria method might be suitable to choose the optimal values. In practice, following Camarda (2012), the smoothing parameters in (3.12) are chosen to be the optimal smoothing parameters for the fit.
Notice that as in the case of the fit, the extension to predict new values depends on the structure of the data. If we consider scattered data, we set n_p out-of-sample points (z_{+_i}, x_{+_i}) at which we want to predict new y_{p_i} values for $i = 1, ..., n_p$ and $\tilde{\boldsymbol{R}}_+$ is a diagonal matrix with the first n values equal to 1 and the last n_p values equal to infinity. Everything else is independent of the data structure.

In the next section, we focus on predictions when just one covariate is extended. In this particular case it is possible to obtain expressions that link the coefficients used in the fit with the ones used in the prediction. This is not possible when we extend the two covariates because of the structure introduced by the Kronecker products.

Prediction of a single covariate

As it is shown in the previous chapter, in one dimension the predicted values depend critically on the order of the penalty. However, once the observed values were fitted, the number of knots, the degree of the P-spline and the smoothing parameter do not have a huge influence on the predicted values. In this section, we show that this is not the case when we work in two dimensions.

In the framework of model (3.1), given a vector of $n_{\boldsymbol{z}} \times n_{\boldsymbol{x}}$ observations $\boldsymbol{y}$ of the response variable, suppose that we want to predict at a new set of values for $\boldsymbol{x}$, $\boldsymbol{x}_p$, therefore, we have $n_p = n_{\boldsymbol{z}} \times n_{\boldsymbol{x}_p}$ new values $\boldsymbol{y}_p$ at $\boldsymbol{z}$ and $\boldsymbol{x}_p$, i.e. we extend just one of the two covariables. Following Currie et al. (2004), we fit and predict the model simultaneously, i.e., we consider the following extended model:

$$\boldsymbol{y}_+ = \boldsymbol{B}_+\boldsymbol{\theta}_+ + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+) \tag{3.15}$$

where $\boldsymbol{y}_+ = (\boldsymbol{y}', \boldsymbol{y}_p')'$, with $\boldsymbol{y}$ the observed values and $\boldsymbol{y}_p$ arbitrary values, and $\boldsymbol{R}_+ = \sigma^2_{\boldsymbol{\epsilon}} \tilde{\boldsymbol{R}}_+$, with $\tilde{\boldsymbol{R}}_+ = \tilde{\boldsymbol{R}}_{\boldsymbol{x}_+} \otimes \tilde{\boldsymbol{R}}_{\boldsymbol{z}}$, and $\tilde{\boldsymbol{R}}_{\boldsymbol{x}_+}$ and $\tilde{\boldsymbol{R}}_{\boldsymbol{z}}$ diagonal matrices of dimensions $n_{\boldsymbol{x}_+} \times n_{\boldsymbol{x}_+}$ and $n_{\boldsymbol{z}} \times n_{\boldsymbol{z}}$, respectively, with infinity entries if the data is to be predicted and 1 if the data is observed, notice that since we are not extending the variable $\boldsymbol{z}$, $\tilde{\boldsymbol{R}}_{\boldsymbol{z}}$ is an identity matrix. In this case, the extended basis is:

$$\boldsymbol{B}_+ = \boldsymbol{B}_{\boldsymbol{x}_+} \otimes \boldsymbol{B}_{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{B}_{\boldsymbol{x}} & \boldsymbol{O} \\ \boldsymbol{B}_{\boldsymbol{x}_{(1)}} & \boldsymbol{B}_{\boldsymbol{x}_{(2)}} \end{bmatrix} \otimes \boldsymbol{B}_{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{B}_{\boldsymbol{x}} \otimes \boldsymbol{B}_{\boldsymbol{z}} & \boldsymbol{O} \\ \boldsymbol{B}_{\boldsymbol{x}_{(1)}} \otimes \boldsymbol{B}_{\boldsymbol{z}} & \boldsymbol{B}_{\boldsymbol{x}_{(2)}} \otimes \boldsymbol{B}_{\boldsymbol{z}} \end{bmatrix}, \tag{3.16}$$

where $\boldsymbol{B}_{\boldsymbol{x}_+} = \boldsymbol{B}(\boldsymbol{x}_+)$ and $\boldsymbol{B}_{\boldsymbol{z}} = \boldsymbol{B}(\boldsymbol{z})$ are the regression bases with $\boldsymbol{x}_+$ and $\boldsymbol{z}$ the two regressors. The new extended B-spline basis, $\boldsymbol{B}_{\boldsymbol{x}_+}$, is built from a new set of knots that consists of the original knots covering x_i, $i = 1, ..., n_{\boldsymbol{x}}$, and extended to the range of the $n_{\boldsymbol{x}_p}$ values of x_{p_j}, $j = 1, ..., n_{\boldsymbol{x}_p}$, i.e., $\boldsymbol{B}_{\boldsymbol{x}_+}$ is a direct extension of $\boldsymbol{B}_{\boldsymbol{x}}$.
Considering the previous extended model, we minimize the function of $\boldsymbol{\theta}_+$ given in (3.11) with $\boldsymbol{B}_+$ defined in (3.16) and extended penalty matrix:

$$\boldsymbol{P}_+ = \lambda_{\boldsymbol{z}} \boldsymbol{P}^{\boldsymbol{z}}_+ + \lambda_{\boldsymbol{x}} \boldsymbol{P}^{\boldsymbol{x}_+}_+, \tag{3.17}$$

where, since only the covariate $\boldsymbol{x}$ is extended, $\boldsymbol{P}^{\boldsymbol{z}}_+$ is:

$$\begin{aligned} \boldsymbol{P}^{\boldsymbol{z}}_+ &= \boldsymbol{I}_{c_{\boldsymbol{x}_+}} \otimes \boldsymbol{D}_{\boldsymbol{z}}' \boldsymbol{D}_{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{I}_{c_{\boldsymbol{x}}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_{c_{\boldsymbol{x}_p}} \end{bmatrix} \otimes \boldsymbol{D}_{\boldsymbol{z}}' \boldsymbol{D}_{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{I}_{c_{\boldsymbol{x}}} \otimes \boldsymbol{D}_{\boldsymbol{z}}' \boldsymbol{D}_{\boldsymbol{z}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_{c_{\boldsymbol{x}_p}} \otimes \boldsymbol{D}_{\boldsymbol{z}}' \boldsymbol{D}_{\boldsymbol{z}} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{P}^{\boldsymbol{z}}_{+11} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{P}^{\boldsymbol{z}}_{+22} \end{bmatrix}, \end{aligned} \tag{3.18}$$

and $\boldsymbol{P}^{\boldsymbol{x}_+}_+$ as in (3.14), where $\boldsymbol{D}_{\boldsymbol{x}_+}$ and $\boldsymbol{D}_{\boldsymbol{z}}$ are the difference matrices acting on the columns and rows of the matrix formed by the extended vector of coefficients, $\boldsymbol{\Theta}_+$:

$$\boldsymbol{\Theta}_+ = \left[\begin{array}{ccccc|ccc} \boldsymbol{\theta}_{11} & \boldsymbol{\theta}_{12} & \cdots & \boldsymbol{\theta}_{1\ c_{\boldsymbol{x}}-1} & \boldsymbol{\theta}_{1c_{\boldsymbol{x}}} & \boldsymbol{\theta}_{1\ c_{\boldsymbol{x}}+1} & \boldsymbol{\theta}_{1\ c_{\boldsymbol{x}}+2} & \cdots \\ \boldsymbol{\theta}_{21} & \boldsymbol{\theta}_{22} & \cdots & \boldsymbol{\theta}_{2\ c_{\boldsymbol{x}}-1} & \boldsymbol{\theta}_{2c_{\boldsymbol{x}}} & \boldsymbol{\theta}_{2\ c_{\boldsymbol{x}}+1} & \boldsymbol{\theta}_{2\ c_{\boldsymbol{x}}+2} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{\theta}_{c_{\boldsymbol{z}}1} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}2} & \cdots & \boldsymbol{\theta}_{c_{\boldsymbol{z}}\ c_{\boldsymbol{x}}-1} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}c_{\boldsymbol{x}}} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}\ c_{\boldsymbol{x}}+1} & \boldsymbol{\theta}_{c_{\boldsymbol{z}}\ c_{\boldsymbol{x}}+2} & \cdots \end{array}\right].$$

As we have said in the previous section, the methodology depends on the structure of the data. If we consider scattered data and extend just the covariate $\boldsymbol{x}$, both bases have to be extended since they have to have the same number of rows, $\boldsymbol{B}_{\boldsymbol{z}}$ is extended by rows to construct $\boldsymbol{B}^+_{\boldsymbol{z}}$ (built from the same knots that $\boldsymbol{B}_{\boldsymbol{z}}$) and $\boldsymbol{B}_{\boldsymbol{x}_+}$ is extended by columns and rows to cover the range of $\boldsymbol{x}_+$. Therefore, $\boldsymbol{B}^+_{\boldsymbol{z}}$ and $\boldsymbol{B}_{\boldsymbol{x}_+}$ have size $n_+ \times c_{\boldsymbol{z}}$ and $n_+ \times c_{\boldsymbol{x}_+}$. The superscript $(+)$ of $\boldsymbol{B}^+_{\boldsymbol{z}}$ indicates that the basis is extended but the prediction is

not outside the range of the observed values of the covariable $\boldsymbol{z}$. In this case $\boldsymbol{R}_+$ is a diagonal matrix with the first n values equal to 1 and the last $n_{\boldsymbol{x}_p}$ values equal to infinity.

Since we extend just one of the two covariates and penalties are based on differences between adjacent coefficients, the method satisfies certain important properties. These properties are an immediate consequence of the following theorems.

Theorem 3.1. *The coefficients obtained from minimization of (3.11) with extended basis (3.16), extended error covariance matrix (3.15) and extended penalty matrix (3.17) satisfy the following properties:*

I. The first c, $c = c_{\boldsymbol{z}} \times c_{\boldsymbol{x}}$, coefficients of $\hat{\boldsymbol{\theta}}_+$, are:

$$\hat{\boldsymbol{\theta}}_{+_{1,\ldots,c}} = \left(\boldsymbol{B}'\boldsymbol{B} + \lambda_{\boldsymbol{x}}\boldsymbol{P}^{\boldsymbol{x}_+}_{+11} + \lambda_{\boldsymbol{z}}\boldsymbol{P}^{\boldsymbol{z}}_{+11} - \lambda^2_{\boldsymbol{x}}\boldsymbol{P}^{\boldsymbol{x}_+}_{+12}(\lambda_{\boldsymbol{x}}\boldsymbol{P}^{\boldsymbol{x}_+}_{+22} + \lambda_{\boldsymbol{z}}\boldsymbol{P}^{\boldsymbol{z}}_{+22})^{-1}\boldsymbol{P}^{\boldsymbol{x}_+}_{+21}\right)^{-1}\boldsymbol{B}'\boldsymbol{y}, \tag{3.19}$$

where $\boldsymbol{P}^{\boldsymbol{x}_+}_{+11}$, $\boldsymbol{P}^{\boldsymbol{x}_+}_{+12}$, $\boldsymbol{P}^{\boldsymbol{x}_+}_{+21}$ and $\boldsymbol{P}^{\boldsymbol{x}_+}_{+22}$ defined in (3.14) and $\boldsymbol{P}^{\boldsymbol{z}}_{+11}$ and $\boldsymbol{P}^{\boldsymbol{z}}_{+22}$ defined in (3.18).

II. The coefficients for the $n_p = n_{\boldsymbol{z}} \times n_{\boldsymbol{x}_p}$ predicted values are

$$\hat{\boldsymbol{\theta}}_p = -\left(\frac{\lambda_{\boldsymbol{z}}}{\lambda_{\boldsymbol{x}}}\boldsymbol{P}^{\boldsymbol{z}}_{+22} + \boldsymbol{P}^{\boldsymbol{x}_+}_{+22}\right)^{-1}\boldsymbol{P}^{\boldsymbol{x}_+}_{+21}\hat{\boldsymbol{\theta}}_{+_{1,\ldots,c}}, \tag{3.20}$$

where $\boldsymbol{P}^{\boldsymbol{x}_+}_{+22}$, $\boldsymbol{P}^{\boldsymbol{x}_+}_{+21}$ defined in (3.14) and $\boldsymbol{P}^{\boldsymbol{z}}_{+22}$ defined in (3.18).

Proof. Differenciating (3.11) with respect to $\boldsymbol{\theta}_+$ leads to

$$\frac{\partial S}{\partial \boldsymbol{\theta}_+} = -2\boldsymbol{B}'_+\tilde{\boldsymbol{R}}^{-1}_+(\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_+) + 2(\lambda_{\boldsymbol{z}}\boldsymbol{P}^{\boldsymbol{z}}_+ + \lambda_{\boldsymbol{x}}\boldsymbol{P}^{\boldsymbol{x}_+}_+) = 0$$

i.e., the penalized least squares solution is given by:

$$\hat{\boldsymbol{\theta}}_+ = (\boldsymbol{B}'_+\tilde{\boldsymbol{R}}^{-1}_+\boldsymbol{B}_+ + \lambda_{\boldsymbol{z}}\boldsymbol{P}^{\boldsymbol{z}}_+ + \lambda_{\boldsymbol{x}}\boldsymbol{P}^{\boldsymbol{x}_+}_+)^{-1}\boldsymbol{B}'_+\tilde{\boldsymbol{R}}^{-1}_+\boldsymbol{y}_+. \tag{3.21}$$

Let us define $\boldsymbol{C} = (\boldsymbol{B}'_+\tilde{\boldsymbol{R}}^{-1}_+\boldsymbol{B}_+ + \lambda_{\boldsymbol{z}}\boldsymbol{P}^{\boldsymbol{z}}_+ + \lambda_{\boldsymbol{x}}\boldsymbol{P}^{\boldsymbol{x}_+}_+)$ and $\boldsymbol{C}^{-1} = \begin{bmatrix}\boldsymbol{C}^{11} & \boldsymbol{C}^{12}\\ \boldsymbol{C}^{21} & \boldsymbol{C}^{22}\end{bmatrix}$, with this notation and since $\tilde{\boldsymbol{R}}^{-1}_+ = \tilde{\boldsymbol{R}}^{-1}_{\boldsymbol{x}_+} \otimes \tilde{\boldsymbol{R}}^{-1}_{\boldsymbol{z}} = \text{blockdiag}(\boldsymbol{I}, \boldsymbol{O})$, with $\boldsymbol{I}$ an identity matrix of dimension $n_{\boldsymbol{x}}n_{\boldsymbol{z}} \times n_{\boldsymbol{x}}n_{\boldsymbol{z}}$ and $\boldsymbol{O}$ a null matrix of dimension $n_{\boldsymbol{x}_p}n_{\boldsymbol{z}} \times n_{\boldsymbol{x}_p}n_{\boldsymbol{z}}$, (3.21) can be rewritten as

$$\boldsymbol{\theta}_+ = \boldsymbol{C}^{-1}\boldsymbol{B}'_+\tilde{\boldsymbol{R}}^{-1}_+\boldsymbol{y}_+ = \begin{bmatrix}\boldsymbol{C}^{11}\boldsymbol{B}'\boldsymbol{y}\\ \boldsymbol{C}^{21}\boldsymbol{B}'\boldsymbol{y}\end{bmatrix}. \tag{3.22}$$

If $\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix}$, by Theorem 8.5.11 given in Harville (2000) we have that:

$$\boldsymbol{C}^{-1} = \begin{bmatrix} \boldsymbol{K}^{-1} & -\boldsymbol{K}^{-1}\boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1} \\ -\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21}\boldsymbol{K}^{-1} & \boldsymbol{C}_{22}^{-1} + \boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21}\boldsymbol{K}^{-1}\boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1} \end{bmatrix},$$

with $\boldsymbol{K} = \boldsymbol{C}_{11} - \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21}$. Therefore:

$$\boldsymbol{C}^{11} = \boldsymbol{K}^{-1} = \left(\boldsymbol{B}'\boldsymbol{B} + \lambda_{\boldsymbol{x}}\boldsymbol{P}_{+11}^{\boldsymbol{x}_+} + \lambda_{\boldsymbol{z}}\boldsymbol{P}_{+11}^{\boldsymbol{z}} - \lambda_{\boldsymbol{x}}^2\boldsymbol{P}_{+12}^{\boldsymbol{x}_+}(\lambda_{\boldsymbol{x}}\boldsymbol{P}_{+22}^{\boldsymbol{x}_+} + \lambda_{\boldsymbol{z}}\boldsymbol{P}_{+22}^{\boldsymbol{z}})^{-1}\boldsymbol{P}_{+21}^{\boldsymbol{x}_+}\right)^{-1}$$

and

$$\begin{aligned} \boldsymbol{C}^{21} &= -\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21}\boldsymbol{K}^{-1} \\ &= -(\lambda_{\boldsymbol{x}}\boldsymbol{P}_{+22}^{\boldsymbol{x}_+} + \lambda_{\boldsymbol{z}}\boldsymbol{P}_{+22}^{\boldsymbol{z}})^{-1}\lambda_{\boldsymbol{x}}\boldsymbol{P}_{+21}^{\boldsymbol{x}_+}\boldsymbol{C}^{11} \end{aligned}$$

and by (3.22) the coefficients for the fit and for the prediction are given by (3.19) and (3.20), respectively, as we wanted to show.

■

Hence, by the previous theorem, when we predict in two dimensions extending one covariate, the predicted values $\boldsymbol{y}_p$ (obtained by using the new coefficients, $\boldsymbol{\theta}_p$) depend on the ratio $\frac{\lambda_{\boldsymbol{z}}}{\lambda_{\boldsymbol{x}}}$, unless $\lambda_{\boldsymbol{x}} = \lambda_{\boldsymbol{z}}$, obviously. Therefore, while in one dimension we have that, once the data are fitted, the smoothing parameter does not play any role in the prediction, we have found that in two dimensions the smoothing parameters in both directions, $\lambda_{\boldsymbol{x}}$ and $\lambda_{\boldsymbol{z}}$, influence the prediction.

Notice that we have proved that the coefficients that give the fit when the fit and the forecast are obtained simultaneously (3.19) are not the same as the solution we obtain only fitting the data (3.7), property that is verified when we predict in one dimension (Section 2.1.1). Although, in the one dimensional case, the extended penalty is not a direct extension of the penalty used to fit the data, the blocks of the extended penalty are simplified and the fit is maintained. This does not occur in the case of two dimensions, unless the block $\boldsymbol{P}_{+22}^{\boldsymbol{z}}$ in (3.18) is equal to zero (or the ratio between $\frac{\lambda_{\boldsymbol{z}}}{\lambda_{\boldsymbol{x}}} \to 0$), as we will see in the following corollary.

Corollary 3.1 (Theorem 3.1)**.** *If $\boldsymbol{P}_{+22}^{\boldsymbol{z}} = \boldsymbol{O}$ in (3.18), the solution from the minimizing (3.11) verifies:*

1. *The fit remains invariant when out-of-sample prediction is carried out.*

2. *Considering the matrix of coefficients that give the fit, $\hat{\boldsymbol{\Theta}}$, and the matrix of coefficients that give the prediction, $\hat{\boldsymbol{\Theta}}_p$, each row $j = 1, ..., c_{\boldsymbol{z}}$, of the additional matrix of coefficients is a linear combination of the last $q_{\boldsymbol{z}}$ old coefficients of that row ($q_{\boldsymbol{z}}$ is the order of the penalty acting on rows and $c_{\boldsymbol{z}}$ the number of rows of $\boldsymbol{B}_{\boldsymbol{z}}$).*

In particular, $\boldsymbol{P}^{\boldsymbol{z}}_{+22} = \boldsymbol{O}$ if $\boldsymbol{I}_{c_{\boldsymbol{x}_p}} = \boldsymbol{O}$.

For the particular case of penalty orders two and three, i.e. $q_{\boldsymbol{x}} = q_{\boldsymbol{z}} = 2$ and $q_{\boldsymbol{x}} = q_{\boldsymbol{z}} = 3$, the proof is given in Appendix B.1.

As we have proved in the previous corollary, setting $\boldsymbol{I}_{c_{\boldsymbol{x}_p}}$ equal to zero we preserve the fit and everything is analogous to the one dimensional case. In the literature, there are some works in which $\boldsymbol{I}_{c_{\boldsymbol{x}_p}}$ is considered equal to zero, e.g. Ugarte et al. (2012). However, in practice, we do not set $\boldsymbol{I}_{c_{\boldsymbol{x}_p}}$ as a null matrix since we are not imposing the penalty correctly. Furthermore, we could not extend it to the case in which we want out-of-sample prediction in both dimensions. We can not set $\boldsymbol{I}_{c_{\boldsymbol{z}_p}}$ and $\boldsymbol{I}_{c_{\boldsymbol{x}_p}}$ equal to zero in (3.13) and (3.14), since the matrix $\boldsymbol{B}'_{+}\tilde{\boldsymbol{R}}^{-1}_{+}\boldsymbol{B}_{+} + \boldsymbol{P}_{+}$ would be singular.

It is important to notice that, regardless the value of $\boldsymbol{P}^{\boldsymbol{x}_+}_{+22}$, (3.20) is telling us that the new coefficients are determined by the coefficients of the last $q_{\boldsymbol{z}}$ columns of the coefficients that give the fit. It means, we do not need to set $\boldsymbol{P}^{\boldsymbol{x}_+}_{+22}$ equal to zero to know which coefficients determine the prediction. However, if $\boldsymbol{P}^{\boldsymbol{x}_+}_{+22}$ is not zero, we can not stablish the relationship between them (i.e., we can not know how strong is the dependence or its shape (linear, quadratic,...)).

In order to preserve the invariance of the fit, in the next section, we propose the use of constraints to maintain the fit when the fit and the prediction are obtained simultaneously, the restrictions can be used when out-of-sample prediction is carried out only in one dimension or in more dimensions.

3.2.2 Constrained out-of-sample prediction

As we have shown in the previous section, natural extensions of penalty matrices provides changes in the fit. To overcome this problem, and as a possible way to incorporate known information about the prediction we propose to use constrained P-splines. In this section we introduce a method that allow us to impose constant and fixed restrictions and to impose restrictions that depend on the observed data.

Our proposal to impose constraints in the prediction is to obtain the solution of the extended models (3.15) and (3.10) subject to a set of l linear constraints given by the equation

$$\boldsymbol{C}\boldsymbol{\theta}_+ = \boldsymbol{r},$$

where $\boldsymbol{C}$ is a constraint matrix of dimension $l \times c_+$ acting on all coefficients, and $\boldsymbol{r}$ is the restrictions vector of dimension $l \times 1$. It means, we have the following restricted extended regression model:

$$\boldsymbol{y}_+ = \boldsymbol{B}_+\boldsymbol{\theta}^*_+ + \boldsymbol{\epsilon}_+, \qquad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+)$$

subject to $\boldsymbol{C}\boldsymbol{\theta}^*_+ = \boldsymbol{r}$. Depending on whether we are predicting out-of-sample in one or two dimensions we extend $\boldsymbol{y}_+$, $\boldsymbol{B}_+$ and $\boldsymbol{R}_+$ defined as in model (3.15) or as in model (3.10). As a clarification on the notation used throughout this document, notice that the superscript $(^*)$ refers to the use of constraints.

We extend the results of Greene and Seaks (1991) to the case of penalized least squares and obtain the Lagrange formulation of the penalized least squares minimization problem:

$$\mathcal{L}(\boldsymbol{\theta}^*_+, \boldsymbol{\omega}) = (\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}^*_+)'\tilde{\boldsymbol{R}}^{-1}_+(\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}^*_+) + \boldsymbol{\theta}^{*'}_+\boldsymbol{P}_+\boldsymbol{\theta}^*_+ + 2\boldsymbol{\omega}'(\boldsymbol{C}\boldsymbol{\theta}^*_+ - \boldsymbol{r}), \qquad (3.23)$$

where $\tilde{\boldsymbol{R}}_+$ defined as in model (3.15) or as in model (3.10) depending on if we extend one or two covariates, $\boldsymbol{P}_+$ is the extended penalty matrix ((3.17) or (3.12)), $\boldsymbol{\theta}^*_+$ denotes the restricted least squares (RLS) estimator and $\boldsymbol{\omega}$ is a $l \times 1$ vector of Lagrange multipliers. Differentiating (3.23) we find

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}^*_+} &= -2\boldsymbol{B}'_+\tilde{\boldsymbol{R}}^{-1}_+(\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}^*_+) + 2\boldsymbol{P}_+\boldsymbol{\theta}^*_+ + 2\boldsymbol{C}'\boldsymbol{\omega}, \qquad (3.24) \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{\omega}} &= \boldsymbol{C}\boldsymbol{\theta}^*_+ - \boldsymbol{r}. \end{aligned}$$

Writing the system as a partitioned matrix the equation yields

$$\begin{bmatrix} \boldsymbol{B}'_+\tilde{\boldsymbol{R}}^{-1}_+\boldsymbol{B}_+ + \boldsymbol{P}_+ & \boldsymbol{C}' \\ \boldsymbol{C} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}}^*_+ \\ \hat{\boldsymbol{\omega}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}'_+\tilde{\boldsymbol{R}}^{-1}_+\boldsymbol{y}_+ \\ \boldsymbol{r} \end{bmatrix}. \qquad (3.25)$$

$\hat{\boldsymbol{\theta}}^*_+$ and $\hat{\boldsymbol{\omega}}$ can be obtained by solving the previous system or, alternative, by following the steps below.

Setting (3.24) to 0, we obtain:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_+^* &= (\boldsymbol{B}'_+\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{B}_+ + \boldsymbol{P}_+)^{-1}(\boldsymbol{B}'_+\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{y}_+ - \boldsymbol{C}'\boldsymbol{\omega}) \\ &= \hat{\boldsymbol{\theta}}_+ - (\boldsymbol{B}'_+\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{B}_+ + \boldsymbol{P}_+)^{-1}\boldsymbol{C}'\hat{\boldsymbol{\omega}}, \end{aligned} \tag{3.26}$$

where $\hat{\boldsymbol{\theta}}_+ = (\boldsymbol{B}'_+\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{B}_+ + \boldsymbol{P}_+)^{-1}\boldsymbol{B}'_+\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{y}_+$ is the unrestricted penalized least squares estimator.

Since $\boldsymbol{C}\boldsymbol{\theta}_+^* = \boldsymbol{r}$, multiplying quation (3.26) by $\boldsymbol{C}$, we have that $\boldsymbol{C}\boldsymbol{\theta}_+ - \boldsymbol{C}(\boldsymbol{B}'_+\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{B}_+ + \boldsymbol{P}_+)^{-1}\boldsymbol{C}'\boldsymbol{\omega} = \boldsymbol{r}$, i.e.

$$\hat{\boldsymbol{\omega}} = [\boldsymbol{C}(\boldsymbol{B}'_+\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{B}_+ + \boldsymbol{P}_+)^{-1}\boldsymbol{C}']^{-1}(\boldsymbol{C}\boldsymbol{\theta}_+ - \boldsymbol{r}). \tag{3.27}$$

Therefore, the coefficients subject to the restriction, $\hat{\boldsymbol{\theta}}_+^*$, are obtained by computing the vector of Lagrange multipliers (3.27) and substituting in (3.26), i.e. $\hat{\boldsymbol{\theta}}_+^*$ is the unconstrained solution, $\hat{\boldsymbol{\theta}}_+$, plus a multiple of the discrepancy vector.

The constrained fitted and predicted values are

$$\hat{\boldsymbol{y}}_+^* = \boldsymbol{B}_+\hat{\boldsymbol{\theta}}_+^*,$$

defining the matrices $\boldsymbol{A_2} = (\boldsymbol{B}'_+\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{B}_+ + \boldsymbol{P}_+)^{-1}\boldsymbol{C}'[\boldsymbol{C}(\boldsymbol{B}'_+\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{B}_+ + \boldsymbol{P}_+)^{-1}\boldsymbol{C}']^{-1}$ and $\boldsymbol{A_1} = (\boldsymbol{B}'_+\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{B}_+ + \boldsymbol{P}_+)^{-1}\boldsymbol{B}'_+$, $\hat{\boldsymbol{y}}_+^*$ can be written as:

$$\hat{\boldsymbol{y}}_+^* = \boldsymbol{B}_+(\boldsymbol{A_1}\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{y}_+ - \boldsymbol{A_2}\boldsymbol{C}\boldsymbol{A_1}\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{y}_+ + \boldsymbol{A_2}\boldsymbol{r}).$$

The variance of $\boldsymbol{y}_+^*$ depends on the following set of restrictions:

a) If the restrictions are constant and fixed, i.e. $\boldsymbol{r}$ is constant and does not depend on the data, the variance is:

$$\mathrm{Var}[\hat{\boldsymbol{y}}_+^*] = \sigma_\epsilon^2\boldsymbol{B}_+\boldsymbol{A_3}\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{A_3}'\boldsymbol{B}'_+,$$

with $\boldsymbol{A_3} = \boldsymbol{A_1} - \boldsymbol{A_2}\boldsymbol{C}\boldsymbol{A_1}$.

b) If the restrictions depend on the data, we have to take into account the variability of $\boldsymbol{r}$. For instance, if the restriction is that the fit has to be maintained, $\boldsymbol{r} = \hat{\boldsymbol{\theta}} = (\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{B}'\boldsymbol{y}$, therefore the variance is:

$$\mathrm{Var}[\hat{\boldsymbol{y}}_+^*] = \sigma_\epsilon^2\boldsymbol{B}_+\boldsymbol{A_4}\tilde{\boldsymbol{R}}_+^{-1}\boldsymbol{A_4}'\boldsymbol{B}'_+,$$

with $\boldsymbol{A_4} = \boldsymbol{A_1} - \boldsymbol{A_2}\boldsymbol{C}\boldsymbol{A_1} + \boldsymbol{A_2}\text{blockdiag}((\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{P})^{-1}, \boldsymbol{O})$, with $\boldsymbol{O}$ a null matrix of dimension $c_p \times n_p$, $c_p = c_{\boldsymbol{z}}c_{\boldsymbol{x}_p} + c_{\boldsymbol{z}_p}c_{\boldsymbol{x}} + c_{\boldsymbol{z}_p}c_{\boldsymbol{x}_p}$ the number of new coefficients, and n_p the number of new observations.

Illustration

Let us explain how the restriction on the fit can be imposed in practice. Suppose that we just carry out out-of-sample prediction in one of the two covariates, that the coefficients matrix from the fit has dimension 4×3, and that the coefficients matrix that gives the fit and the prediction has dimension 4×5 i.e.,

$$\hat{\boldsymbol{\Theta}} = \begin{bmatrix} \hat{\theta}_1 & \hat{\theta}_5 & \hat{\theta}_9 \\ \hat{\theta}_2 & \hat{\theta}_6 & \hat{\theta}_{10} \\ \hat{\theta}_3 & \hat{\theta}_7 & \hat{\theta}_{11} \\ \hat{\theta}_4 & \hat{\theta}_8 & \hat{\theta}_{12} \end{bmatrix}, \qquad \boldsymbol{\Theta}_+^* = \begin{bmatrix} \theta_1 & \theta_5 & \theta_9 & \theta_{13} & \theta_{17} \\ \theta_2 & \theta_6 & \theta_{10} & \theta_{14} & \theta_{18} \\ \theta_3 & \theta_7 & \theta_{11} & \theta_{15} & \theta_{19} \\ \theta_4 & \theta_8 & \theta_{12} & \theta_{16} & \theta_{20} \end{bmatrix},$$

where in $\boldsymbol{\Theta}_+^*$ the coefficients that determine the fit are in red and the coefficients that determine the forecast in blue. If we impose the restriction the fit has to be maintained, we define the restriction equation

$$\boldsymbol{C}\boldsymbol{\theta}_+^* = \boldsymbol{r},$$

where $\boldsymbol{\theta}_+^* = \text{Vec}(\boldsymbol{\Theta}_+^*)$, $\boldsymbol{C} = [\boldsymbol{I}_{12\times 12} \mid \boldsymbol{O}_{12\times 8}]$ ($\boldsymbol{I}_{12\times 12}$ an identity matrix of dimension 12 and $\boldsymbol{O}_{12\times 8}$ a zero matrix of dimension 12×8) and $\boldsymbol{r} = \hat{\boldsymbol{\theta}} = \text{vec}(\hat{\boldsymbol{\Theta}})$.

On the other hand, if we extend the two covariates and the coefficients matrices for the fit and for the fit and the prediction are, respectively:

$$\hat{\boldsymbol{\Theta}} = \begin{bmatrix} \hat{\theta}_1 & \hat{\theta}_5 & \hat{\theta}_9 \\ \hat{\theta}_2 & \hat{\theta}_6 & \hat{\theta}_{10} \\ \hat{\theta}_3 & \hat{\theta}_7 & \hat{\theta}_{11} \\ \hat{\theta}_4 & \hat{\theta}_8 & \hat{\theta}_{12} \end{bmatrix}, \qquad \boldsymbol{\Theta}_+^* = \begin{bmatrix} \theta_1 & \theta_7 & \theta_{13} & \theta_{19} & \theta_{25} \\ \theta_2 & \theta_8 & \theta_{14} & \theta_{20} & \theta_{26} \\ \theta_3 & \theta_9 & \theta_{15} & \theta_{21} & \theta_{27} \\ \theta_4 & \theta_{10} & \theta_{16} & \theta_{22} & \theta_{28} \\ \theta_5 & \theta_{11} & \theta_{17} & \theta_{23} & \theta_{29} \\ \theta_6 & \theta_{12} & \theta_{18} & \theta_{24} & \theta_{30} \end{bmatrix},$$

i.e., $c_{\boldsymbol{z}} = 4$, $c_{\boldsymbol{z}_p} = 2$, $c_{\boldsymbol{x}} = 3$ and $c_{\boldsymbol{x}_p} = 2$. To impose the restriction the fit has to be maintained, we define the restriction equation

$$\boldsymbol{C}\boldsymbol{\theta}_+^* = \boldsymbol{r},$$

where $\boldsymbol{\theta}_+^* = \text{Vec}(\boldsymbol{\Theta}_+^*)$, $\boldsymbol{C} = \text{blockdiag}(\boldsymbol{I}_{4\times4}, [\boldsymbol{O}_{4\times2} \mid \boldsymbol{I}_{4\times4}], [\boldsymbol{O}_{4\times2} \mid \boldsymbol{I}_{4\times4}])$ ($\boldsymbol{I}_{4\times4}$ is an identity matrix of dimension 4 and $\boldsymbol{O}_{4\times2}$ is a zero matrix of dimension 4×2) and $\boldsymbol{r} = \hat{\boldsymbol{\theta}} = \text{vec}(\hat{\boldsymbol{\Theta}})$.

In general, regardless of the number of variables that we extend whenever $c_{\boldsymbol{x}_p} \geq c_{\boldsymbol{z}_p}$, if the restriction is the fit has to be maintained, $\boldsymbol{C}$ is a block diagonal matrix with the first block an identity matrix of dimension $c_{\boldsymbol{z}} \times c_{\boldsymbol{z}}$ and $c_{\boldsymbol{x}_p}$ blocks equal to $[\boldsymbol{O}_{c_{\boldsymbol{z}}\times c_{\boldsymbol{z}_p}} \mid \boldsymbol{I}_{c_{\boldsymbol{z}}\times c_{\boldsymbol{z}}}]$, i.e.

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{I}_{c_{\boldsymbol{z}}\times c_{\boldsymbol{z}}} & & & & \\ & [\boldsymbol{O}_{c_{\boldsymbol{z}}\times c_{\boldsymbol{z}_p}} \mid \boldsymbol{I}_{c_{\boldsymbol{z}}\times c_{\boldsymbol{z}}}] & & & \\ & & [\boldsymbol{O}_{c_{\boldsymbol{z}}\times c_{\boldsymbol{z}_p}} \mid \boldsymbol{I}_{c_{\boldsymbol{z}}\times c_{\boldsymbol{z}}}] & & \\ & & & [\boldsymbol{O}_{c_{\boldsymbol{z}}\times c_{\boldsymbol{z}_p}} \mid \boldsymbol{I}_{c_{\boldsymbol{z}}\times c_{\boldsymbol{z}}}] & \\ & & & & \ddots \end{bmatrix},$$

and $\boldsymbol{r} = \text{vec}(\hat{\boldsymbol{\Theta}})$.

3.2.3 Prediction of mortality data

As in the previous chapter, for the simple purpose of illustrating the proposed methodology, we use a data set on the log mortality rates of US male population considering the log mortality rates as normal data. We use data from the Human Mortality Database (2018), from ages 0 to 110+ over the period 1960-2014, forecasting up to 2050, i.e., we carry out out-of-sample prediction in one of the two covariates, the years.

The Lee-Carter method (Lee and Carter, 1992) is one of the most common methods used for estimating and forecasting mortality data, however this method produces unwanted crossover of forecasted mortality (higher mortality rates for younger ages than for older ages). The original Lee-Carter model is:

$$log(m_{x,y}) = \alpha_x + \beta_x k_y + \epsilon_y,$$

where $m_{x,y}$ is the central rate of morality at age x in year y and α_x, β_x and k_y are parameters to be estimated, and ϵ_y is the error term with mean zero and variance $\sigma_{\boldsymbol{\epsilon}}^2$. This model is fitted to historical data and the resulting estimated k_t's are then modelled and projected as a stochastic time series using standard Box-Jenkins methods. Delwarde et al. (2007) have improved the Lee-Carter model, to avoid the crossover problem, smoothing through penalized splines the estimated β_x's. Other work available in the literature that solves the crossover problem without imposing restrictions is proposed in Currie (2013).

In order to compare our method with the solution given by Delwarde et al. (2007) we have obtained the fit and the prediction through four different models:

a) Model 1, unrestricted model:

$$\boldsymbol{y}_+ = f(\boldsymbol{age}, \boldsymbol{year}_+) + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+)$$

where $\boldsymbol{y}_+$ is the log mortality rate and $\boldsymbol{R}_+$ is defined as in (3.15).

b) Model 2: The model defined in a) subject to the restriction that the fit is maintained.

c) Model 3: The model defined in a) subject to two restrictions:

- The fit is maintained.
- The structure across ages is preserved. We impose this restriction to avoid crossover for ages, i.e. to avoid higher mortality rates for younger ages than for older ages. To do this, we take the coefficients pattern at the last years and we project it. In order to do that we impose that the difference between the coefficients of every two consecutive projections has to be constant and equal to the difference between the corresponding last coefficients from the fit.

  Let us explain with an example how these two restrictions can be imposed at the same time, suppose that the coefficients matrix from the fit has dimension 4×3, and that the coefficients matrix that gives the fit and the forecast has dimension 4×5 i.e.,

$$\hat{\boldsymbol{\Theta}} = \begin{bmatrix} \hat{\theta}_1 & \hat{\theta}_5 & \hat{\theta}_9 \\ \hat{\theta}_2 & \hat{\theta}_6 & \hat{\theta}_{10} \\ \hat{\theta}_3 & \hat{\theta}_7 & \hat{\theta}_{11} \\ \hat{\theta}_4 & \hat{\theta}_8 & \hat{\theta}_{12} \end{bmatrix}, \qquad \boldsymbol{\Theta}^*_+ = \begin{bmatrix} \theta_1 & \theta_5 & \theta_9 & \theta_{13} & \theta_{17} \\ \theta_2 & \theta_6 & \theta_{10} & \theta_{14} & \theta_{18} \\ \theta_3 & \theta_7 & \theta_{11} & \theta_{15} & \theta_{19} \\ \theta_4 & \theta_8 & \theta_{12} & \theta_{16} & \theta_{20} \end{bmatrix},$$

  in $\boldsymbol{\Theta}^*_+$ the coefficients that determine the fit are in red and the coefficients that determine the forecast in blue. The restriction equation is

$$\boldsymbol{C}\boldsymbol{\theta}^*_+ = \boldsymbol{r},$$

  where $\boldsymbol{\theta}^*_+ = \text{Vec}(\boldsymbol{\Theta}^*_+)$, $\boldsymbol{C} = \text{blockdiag}(\boldsymbol{I}_{12\times 12}, \boldsymbol{U}, \boldsymbol{U})$ with $\boldsymbol{U} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$

$$\text{and } \boldsymbol{r} = \begin{bmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\theta}_9 - \hat{\theta}_{10} \\ \hat{\theta}_{10} - \hat{\theta}_{11} \\ \hat{\theta}_{11} - \hat{\theta}_{12} \\ \hat{\theta}_9 - \hat{\theta}_{10} \\ \hat{\theta}_{10} - \hat{\theta}_{11} \\ \hat{\theta}_{11} - \hat{\theta}_{12} \end{bmatrix} \text{ with } \hat{\boldsymbol{\theta}} = \text{Vec}(\hat{\boldsymbol{\Theta}}),$$

i.e., we are imposing that the coefficients that determine the fit have to be the ones we obtain when only fitting the data, and that the difference between the coefficients that determine the forecast of two consecutive rows has to be equal to the difference between the last coefficients from the fit of those rows.

d) Model 4, is the one given in Delwarde et al. (2007).

Figure 3.1 shows the fit and the forecast obtained with model 1 (top left panel), model 2 (top right panel), model 3 (bottom left panel) and model 4 (bottom right panel).

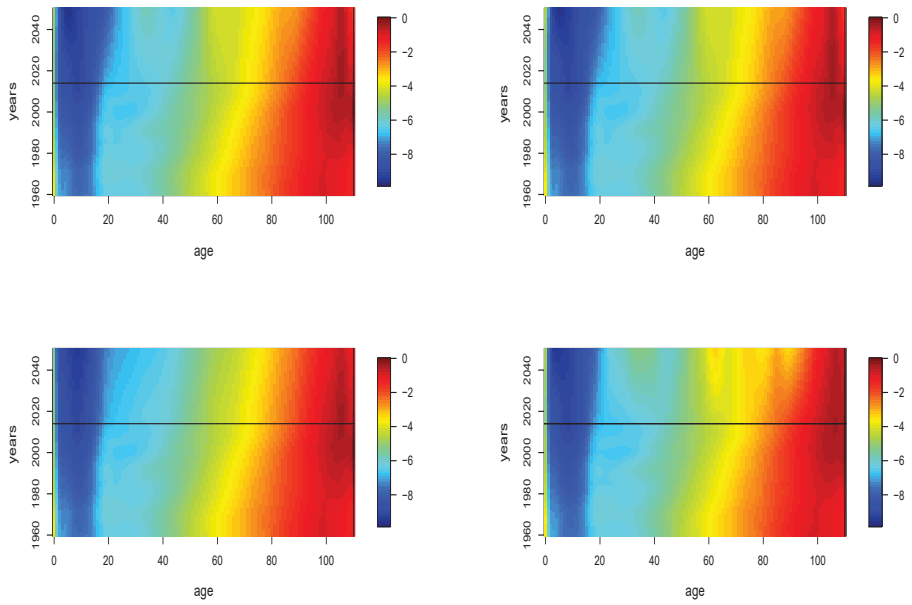


Figure 3.1: Fit and forecast of a data set on the log mortality rates of US males aged 0-110+ over the period 1960-2014, through model 1 (top left panel), model 2 (top right panel), model 3 (bottom left panel) and model 4 (bottom right panel). The horizontal line indicates the year from which we predict (2014).

To highlight the differences between the results, we have selected five ages: 20, 40, 60 and 80, in Figure 3.2 we show the fit and the forecast for those ages obtained through model

1 (red line), model 2 (green line), model 3 (blue line) and model 4 (orange line). The fit provided by the first three models (models 1, 2 and 3) is almost the same. However, the fit given by model 4 is quite different and worse than the others for ages 20 and 40. This could be expected since the Lee-Carter model is not flexible enough to capture the increase in mortality in early ages during the 80's.

The predictions with model 1 and 2 are almost identical (in Figure 3.2 we can hardly appreciate the red line because it is below the green line). Despite giving very similar results in the fit, model 3 provides quite different results in the forecast, as it was expected. For age 60, model 1 and model 2 provide an increase in the log mortality rate for the period $2020 - 2050$ since they are forecasting the incrementing trend in mortality between $2010 - 2016$. This is not consistent with what one would expect. In the case of model 4 forecasts are close to model 3 for ates 40 and 80, but in the case of age 20 it seems to clearly overestimate the mortality. In model 3, the incrementing trend is corrected after the first years to maintain the structure of the adjacent ages, thus avoiding irregular predictions (such as occur at the age of 60 years).

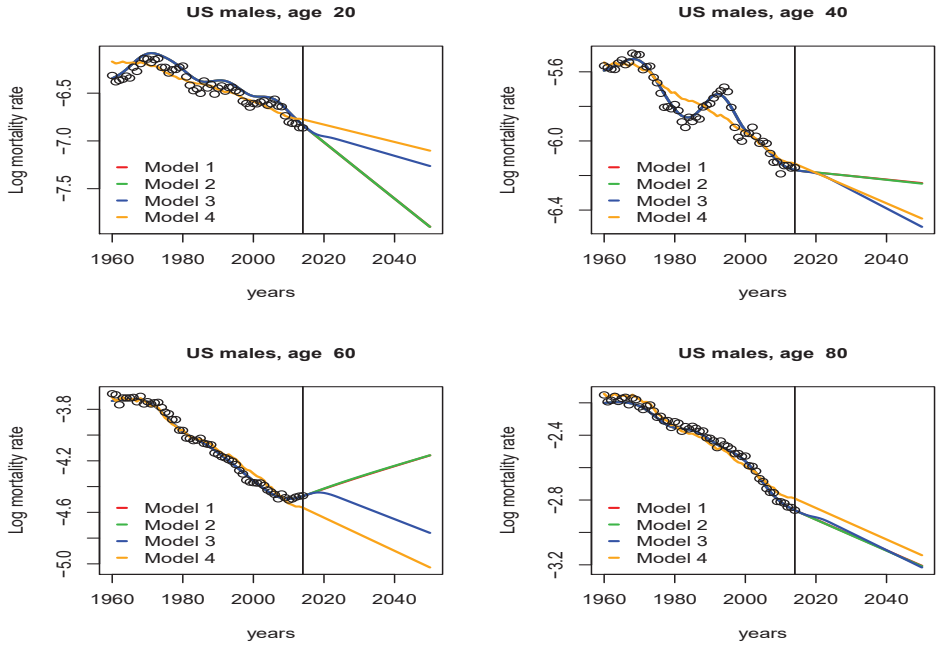


Figure 3.2: Fit and forecast of selected ages: 20, 40, 60 and 80 obtained through model 1 (red line), model 2 (green line), model 3 (blue line) and model 4 (orange line). The vertical line indicates the year from which we predict (2014).

As we have seen, the most realistic results are provided when we impose two restrictions: the fit is maintained and the structure across ages is preserved, i.e. for model 3. If we do

not maintain the coefficients pattern, crossover for ages can happen. We illustrate this fact in Figure 3.3, where we plot the obtained projections with model 2 (green line) and model 3 (blue line) for ages 46 and 47. We can see that the fit is the same for the two models and that the log mortality rates for age 46 are lower than for age 47 in the range of known data. However, in the forecast, for model 2 crossover for ages 46 and 47 occurs and in 2050 the log mortality rate is larger for age 46 than for age 47. This does not occur for model 3, in which case the imposed restriction preserves the structure across ages.

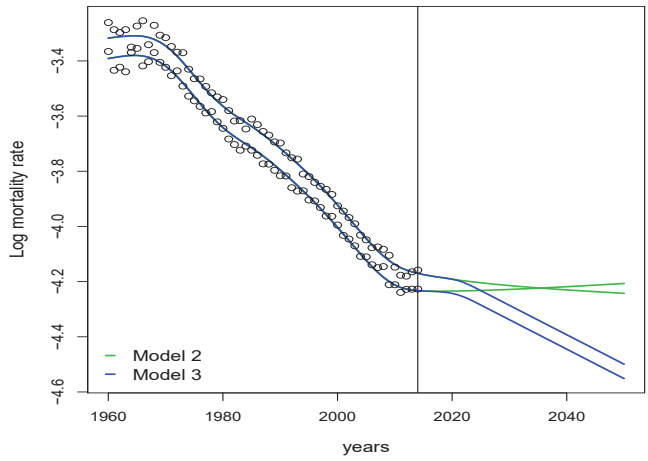


Figure 3.3: Fit and forecast for ages 66 and 67 obtained through model 2 (green line) and model 3 (blue line). Model 3 prevent crossover for ages. The vertical line indicates the year from which we predict (2014).

3.3 Out-of-sample prediction in multidimensional smooth mixed models

Once we have set the general framework, we extend the results presented in Section 3.2.1 to the multidimensional mixed model framework. To reformulate the extended model (3.10) as a mixed model we need to extend the mixed model components to consider the following extended mixed model:

$$\boldsymbol{y}_+ = \boldsymbol{X}_+\boldsymbol{\beta}_{\tilde{+}} + \boldsymbol{Z}_+\boldsymbol{\alpha}_+ + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+), \quad \boldsymbol{\alpha}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}_+), \tag{3.28}$$

The subscript of $\boldsymbol{\beta}_{\tilde{+}}$ is $(\tilde{+})$ and is not $(+)$ to indicate that the fixed effects in the extended model (3.28) are not the same as the fixed effects we obtain only fitting the data, however both fixed effects have the same dimension. The variance matrix of the error,

$\boldsymbol{R}_+$ is defined as in (3.10).

Once we have the extended model matrices, $\boldsymbol{X}_+$ and $\boldsymbol{Z}_+$ and the extended covariance matrix $\boldsymbol{G}_+$, the fit and the forecast are obtained solving the extended mixed model equations of Henderson (1975):

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_{\tilde{+}} \\ \hat{\boldsymbol{\alpha}}_+ \end{bmatrix} = \boldsymbol{L}_+^{-1} \begin{bmatrix} \boldsymbol{X}'_+ \boldsymbol{R}_+^{-1} \\ \boldsymbol{Z}'_+ \boldsymbol{R}_+^{-1} \end{bmatrix} \boldsymbol{y}_+, \tag{3.29}$$

where $\boldsymbol{y}_+$ is defined as in (3.10) and matrix $\boldsymbol{L}_+$ equals

$$\boldsymbol{L}_+ = \begin{bmatrix} \boldsymbol{X}'_+ \boldsymbol{R}_+^{-1} \boldsymbol{X}_+ & \boldsymbol{X}'_+ \boldsymbol{R}_+^{-1} \boldsymbol{Z}_+ \\ \boldsymbol{Z}'_+ \boldsymbol{R}_+^{-1} \boldsymbol{X}_+ & \boldsymbol{Z}'_+ \boldsymbol{R}_+^{-1} \boldsymbol{Z}_+ + \boldsymbol{G}_+^{-1} \end{bmatrix}.$$

Since $\hat{\boldsymbol{y}}_+ = [\boldsymbol{X}_+ \mid \boldsymbol{Z}_+] \begin{bmatrix} \hat{\boldsymbol{\beta}}_{\tilde{+}} \\ \hat{\boldsymbol{\alpha}}_+ \end{bmatrix}$, its variance is:

$$\text{Var}[\hat{\boldsymbol{y}}_+] = [\boldsymbol{X}_+ \mid \boldsymbol{Z}_+] \boldsymbol{L}_+^{-1} \begin{bmatrix} \boldsymbol{X}'_+ \\ \boldsymbol{Z}'_+ \end{bmatrix}.$$

The variance components can be estimated by maximizing the extended residual maximum log-likelihood (REML):

$$-\frac{1}{2}\log|\boldsymbol{V}_+| - \frac{1}{2}\log|\boldsymbol{X}'_+ \boldsymbol{V}_+^{-1} \boldsymbol{X}_+| - \frac{1}{2}(\boldsymbol{y}_+ - \boldsymbol{X}_+ \boldsymbol{\beta}_{\tilde{+}})' \boldsymbol{V}_+^{-1} (\boldsymbol{y}_+ - \boldsymbol{X}_+ \boldsymbol{\beta}_{\tilde{+}}), \tag{3.30}$$

where $\boldsymbol{V}_+ = \boldsymbol{R}_+ + \boldsymbol{Z}_+ \boldsymbol{G}_+ \boldsymbol{Z}'_+$.

To obtain the extended mixed model components we need to define an extended transformation matrix. The natural extension of the transformation matrix (3.8) is to consider the SVD decompositions of the extended difference matrices, i.e. of $\boldsymbol{D}'_{\boldsymbol{z}_+} \boldsymbol{D}_{\boldsymbol{z}_+}$ and of $\boldsymbol{D}'_{\boldsymbol{x}_+} \boldsymbol{D}_{\boldsymbol{x}_+}$, but we have to take into account that the extended transformation built from these singular value decompositions does not provide direct extensions of the mixed model matrices from the fit, $\boldsymbol{X}$ and $\boldsymbol{Z}$. This is not a problem, unless we want to impose restrictions based on the original fit. In this case the fixed effects estimated from the extended model have to be the same as the fixed effects that determine the fit, i.e. $\hat{\boldsymbol{\beta}}_{\tilde{+}} = \hat{\boldsymbol{\beta}}$, and the random effects estimated in the extended model have to be a direct extension of the random effects that determine the fit, i.e. $\hat{\boldsymbol{\alpha}}_+$ has to contain the values of $\hat{\boldsymbol{\alpha}}$. Furthermore, $\hat{\boldsymbol{\beta}}_{\tilde{+}}$ and $\hat{\boldsymbol{\alpha}}_+$ have to be multiplied by model matrices that are direct extensions of the model matrices that determine the fit, $\boldsymbol{X}$ and $\boldsymbol{Z}$.

The natural extended transformation matrix, $\boldsymbol{\Omega}_+$, will not yield the matrices mentioned above, it will return the extended fixed and random effects matrices, $\boldsymbol{X}_+$ and $\boldsymbol{Z}_+$, that are not a direct extension of the model matrices used to obtain the fit, $\boldsymbol{X}$ and $\boldsymbol{Z}$. There are two options that allow us to solve this problem:

- Define the constraint matrix $\boldsymbol{C}$ in the P-spline model framework and reparameterize it to obtain the restrictions matrix for the extended mixed model, $\boldsymbol{C}_{\text{MM}}$, i.e., define $\boldsymbol{C}$ and compute $\boldsymbol{C}_{\text{MM}} = \boldsymbol{C}\boldsymbol{\Omega}_+$.

- Due to identifiability problems the previous proposal can not be used always, as we will see in the next chapter. Therefore, we define an extended transformation matrix $\boldsymbol{\Omega}_+^*$ that allow us to obtain extended fixed and random effects matrices that are a direct extension of the mixed model matrices used to obtain the fit.

The first option is straightforward and can be carried out by using any extended transformation matrix $\boldsymbol{\Omega}_+$. However, to implement the second option we define an extended transformation matrix $\boldsymbol{\Omega}_+^*$ that allow us to preserve the model matrices.

We now give the expressions of the extended mixed model components depending on the extended transformation matrix that we use:

- The natural extended transformation $\boldsymbol{\Omega}_+$ based on the SVD of the extended difference matrices.

- An extended transformation matrix $\boldsymbol{\Omega}_+^*$ that preserves the model matrices from the fit.

3.3.1 Natural reparameterization of P-splines as mixed models for out-of-sample prediction

The natural extension of the transformation matrix (3.8) is to consider the SVD decompositions of the extended difference matrices, i.e. $\boldsymbol{D}'_{\boldsymbol{x}_+}\boldsymbol{D}_{\boldsymbol{x}_+} = \boldsymbol{U}_{\boldsymbol{x}_+}\boldsymbol{\Sigma}_{\boldsymbol{x}_+}\boldsymbol{U}'_{\boldsymbol{x}_+}$ and $\boldsymbol{D}'_{\boldsymbol{z}_+}\boldsymbol{D}_{\boldsymbol{z}_+} = \boldsymbol{U}_{\boldsymbol{z}_+}\boldsymbol{\Sigma}_{\boldsymbol{z}_+}\boldsymbol{U}'_{\boldsymbol{z}_+}$, where the matrices $\boldsymbol{U}_i$, for $i = \boldsymbol{z}_+, \boldsymbol{x}_+$, can be splitted in two parts, $\boldsymbol{U}_i = [\boldsymbol{U}_{if} \mid \boldsymbol{U}_{ir}]$, where $\boldsymbol{U}_{if}$ contains the null part (of dimension $c_i \times q_i$) and $\boldsymbol{U}_{ir}$ contains the span or the non-null part of the decomposition (of dimension $c_i \times (c_i - q_i)$), then the extended transformation matrix is:

$$\boldsymbol{\Omega}_+ = [\underbrace{\boldsymbol{U}_{\boldsymbol{x}_+f} \otimes \boldsymbol{U}_{\boldsymbol{z}f}}_{\boldsymbol{\Omega}_{+f}} \mid \underbrace{\boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{U}_{\boldsymbol{z}_+r} \mid \boldsymbol{U}_{\boldsymbol{x}_+f} \otimes \boldsymbol{U}_{\boldsymbol{z}_+r} \mid \boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{U}_{\boldsymbol{z}_+r}}_{\boldsymbol{\Omega}_{+r}}], \tag{3.31}$$

Given the extended transformation matrix in (3.31), the extended mixed model matrices are:

$$\boldsymbol{X}_+ = \boldsymbol{B}_+\boldsymbol{\Omega}_{+f} = \boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+f} \otimes \boldsymbol{B}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+f}, \tag{3.32}$$

and

$$\mathbf{Z}_+ = \mathbf{B}\boldsymbol{\Omega}_{+r} = \left[\boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{B}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+f} \mid \boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+f} \otimes \boldsymbol{B}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r} \mid \boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{B}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r}\right]. \tag{3.33}$$

Moreover, for the transformation matrix $\boldsymbol{\Omega}_+$ given in (3.31) and the penalty matrix defined in (3.12), the mixed model precision matrix is:

$$\boldsymbol{\Omega}'_+\boldsymbol{P}_+\boldsymbol{\Omega}_+ = \begin{bmatrix} \boldsymbol{O}_q & \\ & \boldsymbol{F}_+ \end{bmatrix}, \text{ with } \boldsymbol{F}_+ = \begin{bmatrix} \lambda_{\boldsymbol{x}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_+} \otimes \boldsymbol{I}_{q_{\boldsymbol{z}}} & & \\ & \lambda_{\boldsymbol{z}}\boldsymbol{I}_{q_{\boldsymbol{x}}} \otimes \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}_+} & \\ & & \lambda_{\boldsymbol{x}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_+} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}_+}-q_{\boldsymbol{z}}} + \lambda_{\boldsymbol{x}}\boldsymbol{I}_{c_{\boldsymbol{x}_+}-q_{\boldsymbol{x}}} \otimes \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}_+} \end{bmatrix},$$

where $q = q_{\boldsymbol{z}}q_{\boldsymbol{x}}$ and the matrices $\tilde{\boldsymbol{\Sigma}}_i$ contains the positive eigenvalues of the SVD of $\boldsymbol{D}'_i\boldsymbol{D}_i$, for $i = \boldsymbol{z}_+, \boldsymbol{x}_+$. Then,

$$\boldsymbol{G}_+ = \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{F}^{-1}_+. \tag{3.34}$$

As a particular case, suppose that we extend just one independent covariable, in this case the natural extension of the trasformation matrix (3.8) is

$$\boldsymbol{\Omega}_+ = [\underbrace{\boldsymbol{U}_{\boldsymbol{x}_+f} \otimes \boldsymbol{U}_{\boldsymbol{z}f}}_{\boldsymbol{\Omega}_{+f}} \mid \underbrace{\boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{U}_{\boldsymbol{z}f} \mid \boldsymbol{U}_{\boldsymbol{x}_+f} \otimes \boldsymbol{U}_{\boldsymbol{z}r} \mid \boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{U}_{\boldsymbol{z}r}}_{\boldsymbol{\Omega}_{+r}}], \tag{3.35}$$

which is based on the SVD decompositions $\boldsymbol{D}'_{\boldsymbol{x}_+}\boldsymbol{D}_{\boldsymbol{x}_+} = \boldsymbol{U}_{\boldsymbol{x}_+}\boldsymbol{\Sigma}_{\boldsymbol{x}_+}\boldsymbol{U}'_{\boldsymbol{x}_+}$ and $\boldsymbol{D}'_{\boldsymbol{z}}\boldsymbol{D}_{\boldsymbol{z}} = \boldsymbol{U}_{\boldsymbol{z}}\boldsymbol{\Sigma}_{\boldsymbol{z}}\boldsymbol{U}'_{\boldsymbol{z}}$. For the previous extended transformation matrix, (3.35), the extended mixed model matrices are:

$$\boldsymbol{X}_+ = \boldsymbol{B}_+\boldsymbol{\Omega}_{+f} = \boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+f} \otimes \boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}f}, \tag{3.36}$$

and

$$\boldsymbol{Z}_+ = \boldsymbol{B}\boldsymbol{\Omega}_{+r} = \left[\boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}f} \mid \boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+f} \otimes \boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}r} \mid \boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}r}\right]. \tag{3.37}$$

Moreover, for the penalty matrix defined in (3.17) and transformation matrix given in in (3.35), the mixed model precision matrix is:

$$\boldsymbol{\Omega}'_+\boldsymbol{P}_+\boldsymbol{\Omega}_+ = \begin{bmatrix} \boldsymbol{O}_q & \\ & \boldsymbol{F}_+ \end{bmatrix}, \text{ with } \boldsymbol{F}_+ = \begin{bmatrix} \lambda_{\boldsymbol{x}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_+} \otimes \boldsymbol{I}_{q_{\boldsymbol{z}}} & & \\ & \lambda_{\boldsymbol{z}}\boldsymbol{I}_{q_{\boldsymbol{x}}} \otimes \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} & \\ & & \lambda_{\boldsymbol{x}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_+} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}-q_{\boldsymbol{z}}} + \lambda_{\boldsymbol{x}}\boldsymbol{I}_{c_{\boldsymbol{x}_+}-q_{\boldsymbol{x}}} \otimes \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} \end{bmatrix},$$

where $q = q_{\boldsymbol{z}}q_{\boldsymbol{x}}$ and the matrices $\tilde{\boldsymbol{\Sigma}}_i$ ($i = \boldsymbol{z}, \boldsymbol{x}_+$) were defined above. And,

$$\boldsymbol{G}_+ = \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{F}^{-1}_+. \tag{3.38}$$

Ugarte et al. (2012) also carried out multidimensional out-of-sample prediction when only one covariate is extended, they set $\boldsymbol{I}_{c_{\boldsymbol{x}_p}}$ equal to zero in (3.17) and propose to use an extended transformation matrix that preserves the transformation used to obtain the fit. They consider the extended transformation matrix $\boldsymbol{\Omega}_+$ defined as $\begin{bmatrix} \boldsymbol{\Omega} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{D}^{-1}_{\boldsymbol{x}_+(2)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}} \end{bmatrix}$, the problem is that with the previous extended transformation matrix we would not differentiate between fixed and random effects. The fixed part would also be penalized since the first q ($q = q_{\boldsymbol{x}} q_{\boldsymbol{z}}$) rows and columns of $\boldsymbol{\Omega}'_+ \boldsymbol{P}_+ \boldsymbol{\Omega}_+$ are not zero.

3.3.2 Reparameterization of P-splines as mixed models for coherent prediction

To preserve the model matrices used to obtain the fit and to penalize the extended random part and not the fixed one, our proposal is to define the following extended transformation matrix:

$$\boldsymbol{\Omega}^*_+ = [\underbrace{\boldsymbol{U}^*_{\boldsymbol{x}_+ f} \otimes \boldsymbol{U}^*_{\boldsymbol{z}_+ f}}_{\boldsymbol{\Omega}^*_{+f}} \mid \underbrace{\boldsymbol{U}^*_{\boldsymbol{x}_+ r} \otimes \boldsymbol{U}^*_{\boldsymbol{z}_+ f} \mid \boldsymbol{U}^*_{\boldsymbol{x}_+ f} \otimes \boldsymbol{U}^*_{\boldsymbol{z}_+ r} \mid \boldsymbol{U}^*_{\boldsymbol{x}_+ r} \otimes \boldsymbol{U}^*_{\boldsymbol{z}_+ r}}_{\boldsymbol{\Omega}^*_{+r}}], \tag{3.39}$$

which is obtained by reordering the block matrices of the matrix $[\boldsymbol{U}^*_{\boldsymbol{x}_+ f} \mid \boldsymbol{U}^*_{\boldsymbol{x}_+ r}] \otimes [\boldsymbol{U}^*_{\boldsymbol{z}_+ f} \mid \boldsymbol{U}^*_{\boldsymbol{z}_+ r}]$, where

$$[\boldsymbol{U}^*_{\boldsymbol{z}_+ f} \mid \boldsymbol{U}^*_{\boldsymbol{z}_+ r}] = \left[\begin{array}{l|ll} \boldsymbol{U}_{\boldsymbol{z}f} & \boldsymbol{U}_{\boldsymbol{z}r} & \boldsymbol{O} \\ -\boldsymbol{D}^{-1}_{\boldsymbol{z}(2)} \boldsymbol{D}_{\boldsymbol{z}(1)} \boldsymbol{U}_{\boldsymbol{z}f} & \boldsymbol{O} & \boldsymbol{D}^{-1}_{\boldsymbol{z}(2)} \end{array}\right], \tag{3.40}$$

$$[\boldsymbol{U}^*_{\boldsymbol{x}_+ f} \mid \boldsymbol{U}^*_{\boldsymbol{x}_+ r}] = \left[\begin{array}{l|ll} \boldsymbol{U}_{\boldsymbol{x}f} & \boldsymbol{U}_{\boldsymbol{x}r} & \boldsymbol{O} \\ -\boldsymbol{D}^{-1}_{\boldsymbol{x}(2)} \boldsymbol{D}_{\boldsymbol{x}(1)} \boldsymbol{U}_{\boldsymbol{x}f} & \boldsymbol{O} & \boldsymbol{D}^{-1}_{\boldsymbol{x}(2)} \end{array}\right], \tag{3.41}$$

and $\boldsymbol{U}_{\boldsymbol{z}f}$, $\boldsymbol{U}_{\boldsymbol{z}r}$, $\boldsymbol{U}_{\boldsymbol{x}f}$ and $\boldsymbol{U}_{\boldsymbol{x}r}$ are defined as in (3.8), $\boldsymbol{D}_{\boldsymbol{z}(2)}$ $\boldsymbol{D}_{\boldsymbol{z}(1)}$ are blocks of the extended difference matrix $\boldsymbol{D}_{\boldsymbol{z}_+}$ (see (3.18)) and $\boldsymbol{D}_{\boldsymbol{x}(2)}$ $\boldsymbol{D}_{\boldsymbol{x}(1)}$ are blocks of the extended difference matrix $\boldsymbol{D}_{\boldsymbol{x}_+}$ (see (3.14)). Notice that this definition of $\boldsymbol{U}^*_{if}$ for $i = \boldsymbol{x}_+, \boldsymbol{z}_+$, verifies $\boldsymbol{D}_i \boldsymbol{U}^*_{if} = \boldsymbol{O}$, i.e., the fixed part is not penalized. However, the previous transformation is not orthogonal and therefore it does not allow us to rewrite the mixed model as a P-spline model.

Then, given the transformation matrix in (3.39) and denoting the matrices $\boldsymbol{X}^*_{\boldsymbol{z}_+} = \boldsymbol{B}_{\boldsymbol{z}_+} \boldsymbol{U}^*_{\boldsymbol{z}_+ f}$, $\boldsymbol{Z}^*_{\boldsymbol{z}_+} = \boldsymbol{B}_{\boldsymbol{z}_+} \boldsymbol{U}^*_{\boldsymbol{z}_+ r}$, $\boldsymbol{X}^*_{\boldsymbol{x}_+} = \boldsymbol{B}_{\boldsymbol{x}_+} \boldsymbol{U}^*_{\boldsymbol{x}_+ f}$ and $\boldsymbol{Z}^*_{\boldsymbol{x}_+} = \boldsymbol{B}_{\boldsymbol{x}_+} \boldsymbol{U}^*_{\boldsymbol{x}_+ r}$, the mixed

model matrices for the two dimensional case are obtained as:

$$\begin{aligned}\boldsymbol{X}_{+}^{*} &= \boldsymbol{X}_{\boldsymbol{x}_{+}}^{*} \otimes \boldsymbol{X}_{\boldsymbol{z}_{+}}^{*},\\ \boldsymbol{Z}_{+}^{*} &= [\boldsymbol{Z}_{\boldsymbol{x}_{+}} \otimes \boldsymbol{X}_{\boldsymbol{z}_{+}}^{*} \mid \boldsymbol{X}_{\boldsymbol{x}_{+}}^{*} \otimes \boldsymbol{Z}_{\boldsymbol{z}_{+}}^{*} \mid \boldsymbol{Z}_{\boldsymbol{x}_{+}}^{*} \otimes \boldsymbol{Z}_{\boldsymbol{z}_{+}}^{*}].\end{aligned}$$

Notice that $\boldsymbol{X}_{i_{+}}^{*}$ and $\boldsymbol{Z}_{i_{+}}^{*}$, for $i = \boldsymbol{z}, \boldsymbol{x}$, are direct extensions of $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$, i.e., they have the following form:

$$\boldsymbol{X}_{i_{+}}^{*} = \begin{bmatrix}\boldsymbol{X}_i\\ \boldsymbol{X}_{i_p}\end{bmatrix},\ \boldsymbol{Z}_{i_{+}}^{*} = \begin{bmatrix}\boldsymbol{Z}_i & \boldsymbol{O}\\ \boldsymbol{Z}_{i(1)} & \boldsymbol{Z}_{i(2)}\end{bmatrix}.$$

Therefore, $\boldsymbol{X}_{+}^{*}$ and $\boldsymbol{Z}_{+}^{*}$ are also direct extensions of $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively, they are:

$$\begin{aligned}\boldsymbol{X}_{+}^{*} &= \begin{bmatrix}\boldsymbol{X}_{\boldsymbol{x}}\\ \boldsymbol{X}_{\boldsymbol{x}_p}\end{bmatrix} \otimes \begin{bmatrix}\boldsymbol{X}_{\boldsymbol{z}}\\ \boldsymbol{X}_{\boldsymbol{z}_p}\end{bmatrix},\\ \boldsymbol{Z}_{+}^{*} &= \begin{bmatrix}\boldsymbol{Z}_{\boldsymbol{x}} & \boldsymbol{O}\\ \boldsymbol{Z}_{\boldsymbol{x}(1)} & \boldsymbol{Z}_{\boldsymbol{x}(2)}\end{bmatrix} \otimes \begin{bmatrix}\boldsymbol{Z}_{\boldsymbol{z}} & \boldsymbol{O}\\ \boldsymbol{Z}_{\boldsymbol{z}(1)} & \boldsymbol{Z}_{\boldsymbol{z}(2)}\end{bmatrix}.\end{aligned}$$

The following theorem gives the covariance matrix of the random effects for the transformation matrix given in (3.39) and the extended penalty matrix given in (3.12).

Theorem 3.2. *Given the extended transformation $\boldsymbol{\Omega}_{+}^{*}$ in two dimensions defined in (3.39) and the extended penalty matrix in (3.12). The mixed model block-diagonal precision matrix $\boldsymbol{F}_{+}^{*}$ is*

$$\boldsymbol{F}_{+}^{*} = \begin{bmatrix}\boldsymbol{F}_{+11}^{*} & \boldsymbol{O} & \boldsymbol{F}_{+13}^{*}\\ \boldsymbol{O} & \boldsymbol{F}_{+22}^{*} & \boldsymbol{F}_{+23}^{*}\\ \boldsymbol{F}_{+31}^{*} & \boldsymbol{F}_{+32}^{*} & \boldsymbol{F}_{+33}^{*}\end{bmatrix} \tag{3.42}$$

with

$$\begin{aligned}\boldsymbol{F}_{+11}^{*} &= \lambda_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*'}\boldsymbol{D}_{\boldsymbol{x}_{+}}'\boldsymbol{D}_{\boldsymbol{x}_{+}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*} \otimes \boldsymbol{U}_{\boldsymbol{z}_{+}f}^{*'}\boldsymbol{U}_{\boldsymbol{z}_{+}f}^{*},\\ \boldsymbol{F}_{+13}^{*} &= \lambda_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*'}\boldsymbol{D}_{\boldsymbol{x}_{+}}'\boldsymbol{D}_{\boldsymbol{x}_{+}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*} \otimes \boldsymbol{U}_{\boldsymbol{z}_{+}f}^{*'}\boldsymbol{U}_{\boldsymbol{z}_{+}r}^{*},\\ \boldsymbol{F}_{+22}^{*} &= \lambda_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{x}_{+}f}^{*'}\boldsymbol{U}_{\boldsymbol{x}_{+}f}^{*} \otimes \boldsymbol{U}_{\boldsymbol{z}_{+}r}^{*'}\boldsymbol{D}_{\boldsymbol{z}_{+}}'\boldsymbol{D}_{\boldsymbol{z}_{+}}\boldsymbol{U}_{\boldsymbol{z}_{+}r}^{*},\\ \boldsymbol{F}_{+23}^{*} &= \lambda_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{x}_{+}f}^{*'}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*} \otimes \boldsymbol{U}_{\boldsymbol{z}_{+}r}^{*'}\boldsymbol{D}_{\boldsymbol{z}_{+}}'\boldsymbol{D}_{\boldsymbol{z}_{+}}\boldsymbol{U}_{\boldsymbol{z}_{+}r}^{*},\\ \boldsymbol{F}_{+31}^{*} &= \boldsymbol{F}_{+13}^{*\prime},\\ \boldsymbol{F}_{+32}^{*} &= \boldsymbol{F}_{+23}^{*\prime},\\ \boldsymbol{F}_{+33}^{*} &= \lambda_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}'\boldsymbol{U}_{\boldsymbol{x}_{+}r} \otimes \boldsymbol{U}_{\boldsymbol{z}_{+}r}'\boldsymbol{D}_{\boldsymbol{z}_{+}}'\boldsymbol{D}_{\boldsymbol{z}_{+}}\boldsymbol{U}_{\boldsymbol{z}_{+}r} + \lambda_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}'\boldsymbol{D}_{\boldsymbol{x}_{+}}'\boldsymbol{D}_{\boldsymbol{x}_{+}}\boldsymbol{U}_{\boldsymbol{x}_{+}r} \otimes \boldsymbol{U}_{\boldsymbol{z}_{+}r}'\boldsymbol{U}_{\boldsymbol{z}_{+}r},\end{aligned}$$

and the covariance matrix of the random effects is $\boldsymbol{G}_{+}^{*} = \sigma_{\boldsymbol{\epsilon}}^{2}\boldsymbol{F}^{*-1}$.

The proof of the previous Theorem is shown in Appendix B.2. Notice that, as we have said, the extended transformation matrix $\boldsymbol{\Omega}_+^*$, defined in (3.39), is not orthogonal, and, moreover, the associated variance-covariance matrix of random effects is not diagonal.

If we extend just one of the two covariates, the extended transformation needed to preserve the model matrices is given by:

$$\boldsymbol{\Omega}_+^* = [\underbrace{\boldsymbol{U}_{\boldsymbol{x}_+f}^* \otimes \boldsymbol{U}_{\boldsymbol{z}f}}_{\boldsymbol{\Omega}_{+f}^*} \mid \underbrace{\boldsymbol{U}_{\boldsymbol{x}_+r}^* \otimes \boldsymbol{U}_{\boldsymbol{z}r} \mid \boldsymbol{U}_{\boldsymbol{x}_+f}^* \otimes \boldsymbol{U}_{\boldsymbol{z}r} \mid \boldsymbol{U}_{\boldsymbol{x}_+r}^* \otimes \boldsymbol{U}_{\boldsymbol{z}r}}_{\boldsymbol{\Omega}_{+r}^*}],$$

which is obtained by reordering the block matrices of the matrix $[\boldsymbol{U}_{\boldsymbol{x}_+f}^* \mid \boldsymbol{U}_{\boldsymbol{x}_+r}^*] \otimes [\boldsymbol{U}_{\boldsymbol{z}f} \mid \boldsymbol{U}_{\boldsymbol{z}r}]$, where $[\boldsymbol{U}_{\boldsymbol{z}f} \mid \boldsymbol{U}_{\boldsymbol{z}r}]$ is defined as in (3.8) and $[\boldsymbol{U}_{\boldsymbol{x}_+f}^* \mid \boldsymbol{U}_{\boldsymbol{x}_+r}^*]$ defined as in (3.39).
In this case, the model components are:

- Extended mixed model matrices:

$$\boldsymbol{X}_+^* = \boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+f}^* \otimes \boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}f},$$

  and

$$\boldsymbol{Z}_+^* = \left[\boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}^* \otimes \boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}f} \mid \boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+f}^* \otimes \boldsymbol{B}_{\boldsymbol{z}}\mathbf{U}_{\boldsymbol{z}r} \mid \boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}^* \otimes \boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}r}\right].$$

- Extended random effects covariance matrix $\boldsymbol{G}_+^* = \sigma_{\boldsymbol{\epsilon}}^2 \boldsymbol{F}^{*-1}$, with

$$\boldsymbol{F}_+^* = \begin{bmatrix} \lambda_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*'}\boldsymbol{D}_{\boldsymbol{x}_+}'\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}^* \otimes \boldsymbol{I}_{q_{\boldsymbol{z}}} & & \\ & \lambda_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{x}_+f}^{*'}\boldsymbol{U}_{\boldsymbol{x}_+f}^* \otimes \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} & \lambda_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{x}_+f}^{*'}\boldsymbol{U}_{\boldsymbol{x}_+r}^* \otimes \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} \\ & \lambda_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*'}\boldsymbol{U}_{\boldsymbol{x}_+f}^* \otimes \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} & \lambda_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{x}_+r}'\boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} + \lambda_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_+r}'\boldsymbol{D}_{\boldsymbol{x}_+}'\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}-q_{\boldsymbol{z}}} \end{bmatrix}$$

  where $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$, of dimensions $(c_{\boldsymbol{z}} - q_{\boldsymbol{z}}) \times (c_{\boldsymbol{z}} - q_{\boldsymbol{z}})$ is the diagonal matrix of positive eigenvalues of $\boldsymbol{D}_{\boldsymbol{z}}'\boldsymbol{D}_{\boldsymbol{z}}$.

As a clarification on the variance components estimation procedure, it is important to say that when we impose invariance of the fit, the variance components that we use are the ones obtained in the fit, i.e. we do not compute them maximizing the extended REML. For us this is a coherent argument, otherwise we would be imposing a fit and using different variance components from the ones that provide that fit. Moreover, to compute the variance components maximizing the extended REML we could not benefit from the algorithms that allow us to compute the variance components quickly (such as the SOP algorithm developed by Rodríguez-Álvarez et al. 2018), since the variance-covariance matrix of the random effects obtained through the extended transformation

matrix defined in (3.39) is not diagonal.

3.3.3 Constrained smooth mixed models for coherent out-of-sample prediction

As in the case of 2D P-spline models, constraints need to be imposed in order to ensure coherent fit and prediction. In this section we explain how predictions (subject to the restriction that the fit is kept) are carried out in the context of mixed models. Suppose that we consider the following restricted extended mixed model:

$$\boldsymbol{y}_+^* = \boldsymbol{X}_+\beta_+^* + \boldsymbol{Z}_+\boldsymbol{\alpha}_+^* + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+), \quad \boldsymbol{\alpha}_+^* \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}_+),$$

subject to $\boldsymbol{C}_{\text{MM}} \begin{bmatrix} \beta_+^* \\ \boldsymbol{\alpha}_+^* \end{bmatrix} = \boldsymbol{r}_{\text{MM}}$, where $\boldsymbol{C}_{\text{MM}}$ is a constraint matrix of dimension $l \times c_+$ acting on all coefficients and built analogously to how $\boldsymbol{C}$ is built in Section 3.2.2, and $\boldsymbol{r}_{\text{MM}} = \begin{bmatrix} \hat{\beta} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix}$ is the restrictions vector of dimension $l \times 1$. Notice that we use the superscript $(*)$ to indicate that we are imposing restrictions.

To estimate the restricted parameters we minimize the following constrained penalized likelihood:

$$\begin{aligned} \mathcal{L}(\beta_+^*, \boldsymbol{\alpha}_+^*, \boldsymbol{w}) &= (\boldsymbol{y}_+ - \boldsymbol{X}_+\beta_+^* - \boldsymbol{Z}_+\boldsymbol{\alpha}_+^*)'\boldsymbol{R}_+^{-1}(\boldsymbol{y}_+ - \boldsymbol{X}_+\beta_+^* - \boldsymbol{Z}_+\boldsymbol{\alpha}_+^*) + \boldsymbol{\alpha}_+^{*'}\boldsymbol{G}_+^{-1}\boldsymbol{\alpha}_+^* \\ &\quad +2\boldsymbol{w}'\left(\boldsymbol{C}_{MM}\begin{bmatrix} \beta_+^* \\ \boldsymbol{\alpha}_+^* \end{bmatrix} - \boldsymbol{r}_{MM}\right) \end{aligned} \tag{3.43}$$

Since we have to take derivatives with respect to $\boldsymbol{\beta}_+^*$ and with respect to $\boldsymbol{\alpha}_+^*$, we divide the matrix of constraints into two parts, one associated with the fixed effects and the other one associated with the random effects, $\boldsymbol{C}_{MM} = \left[\boldsymbol{C}_{MM_f} \mid \boldsymbol{C}_{MM_r}\right]$, and rewrite (3.43) as:

$$\begin{aligned} \mathcal{L}(\beta_+^*, \boldsymbol{\alpha}_+^*, \boldsymbol{w}) &= (\boldsymbol{y}_+ - \boldsymbol{X}_+\beta_+^* - \boldsymbol{Z}_+\boldsymbol{\alpha}_+^*)'\boldsymbol{R}_+^{-1}(\boldsymbol{y}_+ - \boldsymbol{X}_+\beta_+^* - \boldsymbol{Z}_+\boldsymbol{\alpha}_+^*) + \boldsymbol{\alpha}_+^{*'}\boldsymbol{G}_+^{-1}\boldsymbol{\alpha}_+^* \\ &\quad +2\boldsymbol{\omega}'(\boldsymbol{C}_{MM_f}\beta_+^* + \boldsymbol{C}_{MM_r}\boldsymbol{\alpha}_+^* - \boldsymbol{r}_{MM}). \end{aligned}$$

The first order conditions yield

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}_+^*} = -2\boldsymbol{X}'_+\boldsymbol{R}_+^{-1}(\boldsymbol{y}_+ - \boldsymbol{X}_+\beta_+^* - \boldsymbol{Z}_+\boldsymbol{\alpha}_+^*) + 2\boldsymbol{C}'_{MM_f}\boldsymbol{\omega} \tag{3.44}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}_+^*} = -2\boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}(\boldsymbol{y}_+ - \boldsymbol{X}_+\beta_+^* - \boldsymbol{Z}_+\boldsymbol{\alpha}_+^*) + 2\boldsymbol{G}_+^{-1}\boldsymbol{\alpha}_+ + 2\boldsymbol{C}'_{MM_r}\boldsymbol{\omega} \tag{3.45}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \boldsymbol{C}_{MM_f}\boldsymbol{\beta}_+^* + \boldsymbol{C}_{MM_r}\boldsymbol{\alpha}_+^* - \boldsymbol{r}_{MM} \tag{3.46}$$

Therefore, we have the system:

$$\begin{bmatrix} \boldsymbol{X}'_+\boldsymbol{R}_+^{-1}\boldsymbol{X}_+ & \boldsymbol{X}'_+\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+ & \boldsymbol{C}'_{MM_f} \\ \boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}\boldsymbol{X}_+ & \boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+ + \boldsymbol{G}_+^{-1} & \boldsymbol{C}'_{MM_r} \\ \boldsymbol{C}_{MM_f} & \boldsymbol{C}_{MM_r} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^*_{\tilde{+}} \\ \boldsymbol{\alpha}^*_+ \\ \boldsymbol{\omega} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'_+\boldsymbol{R}_+^{-1}\boldsymbol{y}_+ \\ \boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}\boldsymbol{y}_+ \\ \boldsymbol{r}_{MM} \end{bmatrix}$$

We can also compute the fixed and random effects and the Lagrangean multipliers following the next steps. By (3.44) and (3.45) we know that

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}^*_{\tilde{+}} \\ \hat{\boldsymbol{\alpha}}^*_+ \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_{\tilde{+}} \\ \hat{\boldsymbol{\alpha}}_+ \end{bmatrix} - \boldsymbol{L}_+^{-1}\boldsymbol{C}'_{MM}\hat{\boldsymbol{w}}, \tag{3.47}$$

where $\boldsymbol{L}_+ = \begin{bmatrix} \boldsymbol{X}'_+\boldsymbol{R}_+^{-1}\boldsymbol{X}^*_+ & \boldsymbol{X}'_+\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+ \\ \boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}\boldsymbol{X}_+ & \boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+ + \boldsymbol{G}_+^{-1} \end{bmatrix}$ and $\begin{bmatrix} \hat{\boldsymbol{\beta}}_{\tilde{+}} \\ \hat{\boldsymbol{\alpha}}_+ \end{bmatrix}$ are the unrestricted penalized least squares estimators, and

$$\hat{\boldsymbol{\omega}} = \left[\boldsymbol{C}_{MM}\boldsymbol{L}_+^{-1}\boldsymbol{C}'_{MM}\right]^{-1}\left[\boldsymbol{C}_{MM}\begin{bmatrix} \boldsymbol{\beta}_{\tilde{+}} \\ \boldsymbol{\alpha}_+ \end{bmatrix} - \boldsymbol{r}_{MM}\right]. \tag{3.48}$$

Therefore, the coefficients subject to the restriction, $\begin{bmatrix} \hat{\boldsymbol{\beta}}^*_{\tilde{+}} \\ \hat{\boldsymbol{\alpha}}^*_+ \end{bmatrix}$, are obtained by computing the vector of Lagrange multipliers (3.48) and substituting in (3.47), i.e. $\begin{bmatrix} \hat{\boldsymbol{\beta}}^*_{\tilde{+}} \\ \hat{\boldsymbol{\alpha}}^*_+ \end{bmatrix}$ is the unconstrained solution, $\begin{bmatrix} \hat{\boldsymbol{\beta}}_{\tilde{+}} \\ \hat{\boldsymbol{\alpha}}_+ \end{bmatrix}$, plus a multiple of the discrepancy vector.

The restricted fitted and predicted values are:

$$\hat{\boldsymbol{y}}^*_+ = [\boldsymbol{X}_+ \mid \boldsymbol{Z}_+]\begin{bmatrix} \hat{\boldsymbol{\beta}}^*_* \\ \hat{\boldsymbol{\alpha}}^*_+ \end{bmatrix},$$

defining the matrices $\boldsymbol{L}_+ = \begin{bmatrix} \boldsymbol{X}'_+\boldsymbol{R}_+^{-1}\boldsymbol{X}_+ & \boldsymbol{X}'_+\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+ \\ \boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}\boldsymbol{X}_+ & \boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+ + \boldsymbol{G}_+^{-1} \end{bmatrix}$, $\boldsymbol{A_1}_{MM} = \boldsymbol{L}_+^{-1}\begin{bmatrix} \boldsymbol{X}_+ \\ \boldsymbol{Z}_+ \end{bmatrix}$ and $\boldsymbol{A_2}_{MM} = \boldsymbol{L}_+^{-1}\boldsymbol{C}'_{MM}[\boldsymbol{C}_{MM}\boldsymbol{L}_+^{-1}\boldsymbol{C}'_{MM}]^{-1}$, $\hat{\boldsymbol{y}}_+$ can be written as:

$$\hat{\boldsymbol{y}}^*_+ = [\boldsymbol{X}_+ \mid \boldsymbol{Z}_+](\boldsymbol{A_1}_{MM}\boldsymbol{R}_+^{-1}\boldsymbol{y}_+ - \boldsymbol{A_2}_{MM}\boldsymbol{C}_{MM}\boldsymbol{A_1}_{MM}\boldsymbol{R}_+^{-1}\boldsymbol{y}_+ + \boldsymbol{A_2}_{MM}\boldsymbol{r}_{MM}).$$

Since $\boldsymbol{r}_{MM} = \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} = \boldsymbol{L}\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1} \end{bmatrix}\boldsymbol{y}$, with $\boldsymbol{L} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix}$, taking into

account the variability of $\boldsymbol{r}_{MM}$ the variance is:

$$\text{Var}[\hat{\boldsymbol{y}}_+^*] = [\boldsymbol{X}_+ \mid \boldsymbol{Z}_+]\boldsymbol{A}_{\boldsymbol{4}MM}\boldsymbol{R}_+^{-1}\boldsymbol{A}'_{\boldsymbol{4}MM}\begin{bmatrix}\boldsymbol{X}'_+\\ \boldsymbol{Z}'_+\end{bmatrix},$$

with $\boldsymbol{A}_{\boldsymbol{4}MM} = \left(\boldsymbol{A}_{\boldsymbol{1}MM} - \boldsymbol{A}_{\boldsymbol{2}MM}\boldsymbol{C}_{MM}\boldsymbol{A}_{\boldsymbol{1}MM} + \boldsymbol{A}_{\boldsymbol{2}MM}\text{blockdiag}\left(\boldsymbol{L}\begin{bmatrix}\boldsymbol{X}'\\ \boldsymbol{Z}'\end{bmatrix}, \boldsymbol{O}\right)\right)$, with $\boldsymbol{O}$ a null matrix of dimension $c_p \times n_p$, c_p the number of new coefficients and n_p the number of new values we predict.

3.4 Summary of the chapter

In this chapter, we have presented a general framework for out-of-sample prediction in smooth additive models with interaction terms. We build our proposal from the method proposed in Currie et al. (2004), and we have extended their approach to the case in which out-of-sample prediction is necessary in both directions of the interaction terms. The method proposed in Currie et al. (2004) deal with out-of-sample predictions as missing values with 0 weights and carry out fit and prediction simultaneously. We have shown that this approach yields different fitted values depending on whether only fitting or fitting and prediction is carried out. To solve this incoherence we propose to maximize the penalized likelihood subject to linear constraints which ensure that the coefficients obtained in the fit of the data remain the same when fit and out-of-sample prediction are carried out simultaneously. To do so we use Lagrange multipliers. This general approach can be used to impose other relevant constraints such as the case of mortality forecasting when structure across ages needs to be preserved. The methodology proposed is extended to the case of smooth mixed models since it will allow us to predict out-of-sample in a wide class of models. We have shown that attention needs to be imposed since the matrices of fixed and random effects for out-of-sample prediction need to be direct extensions of the matrices used in the fit.
The constrained prediction method proposed has also been used in one real data example, one in which mortality rates are forecasted over the years and the importance of imposing constraints on the coefficients to ensure coherent forecast is shown.

Chapter 4

Component-wise prediction with P-spline Smooth-ANOVA models

In this chapter, we extend the previous methodology to the so-called smooth-ANOVA models which allow us to include interaction terms that can be decomposed as a sum of several smooth functions. The structure of this chapter is the following, our proposal to predict new values in a more flexible context is shown in Section 4.2, where we give results on out-of-sample prediction for the Smooth-ANOVA model of Lee and Durbán (2011). In Section 4.2.3 we illustrate the proposed methodology, reanalyzing the data set used in Section 2.3.1 to predict aboveground biomass in Populus trees as a smooth function of height and diameter, but in this case including interaction terms. Finally, in Section 4.3 we examine the performance of the interaction models in comparison to the Smooth-ANOVA models (both models with and without imposing invariance of the fit) through a simulation study.

4.1 P-spline Smooth-ANOVA models

Sometimes, model (3.1) will not be flexible enough and it might force unnecessary complexity (i.e. use too many degrees of freedom to fit the data). In order to add more flexibility and drop unnecessary terms if they are not relevant, Lee and Durbán (2011) propose the use of P-spline Smooth-ANOVA models. This approach decomposes the interaction terms in a similar way as the analysis of variance does. In this section, we show how carry out out-of-sample prediction in this context. First, we briefly review how we can fit a smooth interaction function as a decomposition of smooth functions which are identifiable.

Suppose we have array data and a data vector $\boldsymbol{y}$ of length $n \times 1$, where $n = n_{\boldsymbol{z}} n_{\boldsymbol{x}}$, and the regressors $\boldsymbol{z} = (z_1, z_2, ...z_{n_{\boldsymbol{z}}})'$ and $\boldsymbol{x} = (x_1, x_2, ...x_{n_{\boldsymbol{x}}})'$. Let us consider the following

Smooth-ANOVA model:

$$\boldsymbol{y} = \gamma + f_1(\boldsymbol{z}) + f_2(\boldsymbol{x}) + f_{1,2}(\boldsymbol{z}, \boldsymbol{x}) + \boldsymbol{\epsilon} = \boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}) \tag{4.1}$$

where $\boldsymbol{R} = \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}$, i.e. the errors are independent and identically distributed and the B-spline basis $\boldsymbol{B}$ is defined as:

$$\boldsymbol{B} = [\boldsymbol{1}_n \mid \boldsymbol{1}_{n_{\boldsymbol{x}}} \otimes \boldsymbol{B}_{\boldsymbol{z}} \mid \boldsymbol{B}_{\boldsymbol{x}} \otimes \boldsymbol{1}_{n_{\boldsymbol{z}}} \mid \boldsymbol{B}_{\boldsymbol{x}} \otimes \boldsymbol{B}_{\boldsymbol{z}}],$$

of dimension $n \times (1 + c_{\boldsymbol{z}} + c_{\boldsymbol{x}} + c_{\boldsymbol{z}}c_{\boldsymbol{x}})$, and where $\boldsymbol{1}_{n_{\boldsymbol{z}}}$ and $\boldsymbol{1}_{n_{\boldsymbol{x}}}$ are column vectors of ones of length $n_{\boldsymbol{z}}$ and $n_{\boldsymbol{x}}$ respectively, and the vector of regression coefficients is $\boldsymbol{\theta} = (\gamma, \boldsymbol{\theta}'_{\boldsymbol{z}}, \boldsymbol{\theta}'_{\boldsymbol{x}}, \boldsymbol{\theta}'_{\boldsymbol{s}})'$, where $\boldsymbol{\theta}_{\boldsymbol{z}}$ and $\boldsymbol{\theta}_{\boldsymbol{x}}$ are the vectors of coefficients for the main effects, of dimension $c_{\boldsymbol{z}} \times 1$ and $c_{\boldsymbol{x}} \times 1$, respectively, and $\boldsymbol{\theta}_s$ is the vector of coefficients for the interaction, of dimension $c_{\boldsymbol{z}}c_{\boldsymbol{x}} \times 1$. Therefore, model (4.1) is written as:

$$\boldsymbol{y} = \gamma\boldsymbol{1}_n + (\boldsymbol{1}_{n_{\boldsymbol{x}}} \otimes \boldsymbol{B}_{\boldsymbol{z}})\boldsymbol{\theta}_{\boldsymbol{z}} + (\boldsymbol{B}_{\boldsymbol{x}} \otimes \boldsymbol{1}_{n_{\boldsymbol{z}}})\boldsymbol{\theta}_{\boldsymbol{x}} + (\boldsymbol{B}_{\boldsymbol{x}} \otimes \boldsymbol{B}_{\boldsymbol{z}})\boldsymbol{\theta}_{\boldsymbol{s}}, \tag{4.2}$$

with penalty the following block-diagonal matrix:

$$\boldsymbol{P} = \begin{bmatrix} 0 & & & \\ & \lambda_{\boldsymbol{z}}\boldsymbol{D}'_{\boldsymbol{z}}\boldsymbol{D}_{\boldsymbol{z}} & & \\ & & \lambda_{\boldsymbol{x}}\boldsymbol{D}'_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{x}} & \\ & & & \tau_{\boldsymbol{z}}\boldsymbol{I}_{c_{\boldsymbol{x}}} \otimes \boldsymbol{D}'_{\boldsymbol{z}}\boldsymbol{D}_{\boldsymbol{z}} + \tau_{\boldsymbol{x}}\boldsymbol{D}'_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{x}} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}} \end{bmatrix}, \tag{4.3}$$

of dimension $(1+c_{\boldsymbol{z}}+c_{\boldsymbol{x}}+c_{\boldsymbol{z}}c_{\boldsymbol{x}}) \times (1+c_{\boldsymbol{z}}+c_{\boldsymbol{x}}+c_{\boldsymbol{z}}c_{\boldsymbol{x}})$, where each block corresponds to the penalty over each of the coefficients of the model. Lee (2010) pointed out that the model (4.2) and the penalty (4.3) should be modified in order to preserve identifiability, their proposal is to construct identifiability model bases and penalties reparameterizing the model as a mixed model instead of imposing numerical constraints as other authors have proposed (Wood 2006). They define the following transformation matrix $\boldsymbol{\Omega} = [\boldsymbol{\Omega}_f \mid \boldsymbol{\Omega}_r]$ with dimension $(1 + c_{\boldsymbol{z}} + c_{\boldsymbol{x}} + c_{\boldsymbol{z}}c_{\boldsymbol{x}})$, where:

$$\boldsymbol{\Omega}_f = \begin{bmatrix} 1 & & & \\ & 1 \otimes \boldsymbol{u}^{(2)}_{\boldsymbol{z}f} & & \\ & & \boldsymbol{u}^{(2)}_{\boldsymbol{x}f} \otimes 1 & \\ & & & \boldsymbol{u}^{(2)}_{\boldsymbol{x}f} \otimes \boldsymbol{u}^{(2)}_{\boldsymbol{z}f} \end{bmatrix},$$

$$\boldsymbol{\Omega}_r = \begin{bmatrix} 1 & & & \\ & 1 \otimes \boldsymbol{U}_{zr} & & \\ & & \boldsymbol{U}_{xr} \otimes 1 & \\ & & & \boldsymbol{u}_{xf}^{(2)} \otimes \boldsymbol{U}_{zr} \mid \boldsymbol{U}_{xr} \otimes \boldsymbol{u}_{zf}^{(2)} \mid \boldsymbol{U}_{xr} \otimes \boldsymbol{U}_{zr} \end{bmatrix},$$

and $\boldsymbol{u}_{zf}^{(2)}$ and $\boldsymbol{u}_{xf}^{(2)}$ are the second columns of $\boldsymbol{U}_{zf}$ and $\boldsymbol{U}_{xf}$, respectively, and $\boldsymbol{U}_{if}$ and $\boldsymbol{U}_{ir}$ are the eigenvectors corresponding to the zero values and positive values of the SVD of $\boldsymbol{D}_i'\boldsymbol{D}_i$, respectively, for $i = \boldsymbol{x}, \boldsymbol{z}$.

Given the previous transformation matrix, Lee (2010) shows that the fixed and random effects matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are:

$$\begin{aligned} \boldsymbol{X} &= [\mathbf{1}_n \mid \mathbf{1}_{n_{\boldsymbol{x}}} \otimes \tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}} \otimes \mathbf{1}_{n_{\boldsymbol{z}}} \mid \tilde{\boldsymbol{x}} \otimes \tilde{\boldsymbol{z}}] \\ \boldsymbol{Z} &= [\mathbf{1}_n \mid \mathbf{1}_{n_{\boldsymbol{x}}} \otimes \boldsymbol{Z}_{\boldsymbol{z}} \mid \boldsymbol{Z}_{\boldsymbol{x}} \otimes \mathbf{1}_{n_{\boldsymbol{z}}} \mid \tilde{\boldsymbol{x}} \otimes \boldsymbol{Z}_{\boldsymbol{z}} \mid \boldsymbol{Z}_{\boldsymbol{x}} \otimes \tilde{\boldsymbol{z}} \mid \boldsymbol{Z}_{\boldsymbol{x}} \otimes \boldsymbol{Z}_{\boldsymbol{z}}] \end{aligned} \tag{4.4}$$

where $\tilde{\boldsymbol{z}} = \boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{u}_{zf}^{(2)}$, $\tilde{\boldsymbol{x}} = \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{u}_{xf}^{(2)}$, $\boldsymbol{Z}_{\boldsymbol{z}} = \boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{zr}$ and $\boldsymbol{Z}_{\boldsymbol{x}} = \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{U}_{xr}$. Moreover, the mixed model penalty is:

$$\boldsymbol{F} = \text{blockdiag}(\boldsymbol{F}_{(1)}, \boldsymbol{F}_{(2)}, \boldsymbol{F}_{(1,2)}),$$

where for a second order penalty, it has size $(c_{\boldsymbol{z}}c_{\boldsymbol{x}} - 4) \times (c_{\boldsymbol{z}}c_{\boldsymbol{x}} - 4)$, and with:

$$\begin{aligned} \boldsymbol{F}_{(1)} &= \lambda_{\boldsymbol{z}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}}, \\ \boldsymbol{F}_{(2)} &= \lambda_{\boldsymbol{x}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}, \\ \boldsymbol{F}_{(1,2)} &= \text{blockdiag}(\tau_{\boldsymbol{z}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}}, \tau_{\boldsymbol{x}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}, \tau_{\boldsymbol{z}}\boldsymbol{I}_{c_{\boldsymbol{x}}-2} \otimes \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} + \tau_{\boldsymbol{x}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}-2}), \end{aligned}$$

where $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ are the nonzero eigenvalues of the SVD of $\boldsymbol{D}_{\boldsymbol{z}}'\boldsymbol{D}_{\boldsymbol{z}}$ and $\boldsymbol{D}_{\boldsymbol{x}}'\boldsymbol{D}_{\boldsymbol{x}}$, respectively. With the previous representation of model (4.1), Lee (2010) avoid the identifiability problem removing the column vector of $\mathbf{1}$'s in the random effects matrix (4.4). For more details, see Lee (2010). Then, we will detail our proposal to obtain out-of-sample predictions with S-ANOVA models.

4.2 Out-of-sample prediction with P-spline Smooth-ANOVA models

Although the out-of-sample prediction will be carried out in the context of mixed models, we present the approach in the original P-splines formulation, since the reparameterization needed for the out-of-sample prediction will be based on this formulation. In particular, we need to know the extended penalty matrix in order to calculate the pre-

cision matrix of the random effects as we will see in Sections 4.2.1 and 4.2.2.

In the framework of model (4.1), given a vector of $n_{\boldsymbol{z}} n_{\boldsymbol{x}} \times 1$ observations $\boldsymbol{y}$ of the response variable, suppose that we want to predict $n_p = n_{\boldsymbol{z}} n_{\boldsymbol{x}_p} + n_{\boldsymbol{z}_p} n_{\boldsymbol{x}} + n_{\boldsymbol{z}_p} n_{\boldsymbol{x}_p}$ new values at $(\boldsymbol{z}, \boldsymbol{x}_p)$, $(\boldsymbol{z}_p, \boldsymbol{x})$ and $(\boldsymbol{z}_p, \boldsymbol{x}_p)$. I.e., the matrix $\boldsymbol{Y}_+$ of observed and predicted values can be arrenged as in (3.9). For this case we consider the following extended Smooth-ANOVA model:

$$\boldsymbol{y}_+ = \gamma + f_1(\boldsymbol{z}_+) + f_2(\boldsymbol{x}_+) + f_{1,2}(\boldsymbol{z}_+, \boldsymbol{x}_+) + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+) \tag{4.5}$$

where $\boldsymbol{R}_+$ is defined as in (3.10) and we assume:

$$\boldsymbol{y}_+ = \boldsymbol{B}_+ \boldsymbol{\theta}_+ + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+)$$

where the extended B-spline basis $\boldsymbol{B}_+$ is defined as:

$$\boldsymbol{B}_+ = [\boldsymbol{1}_{n_+} \mid \boldsymbol{1}_{n_{\boldsymbol{x}_+}} \otimes \boldsymbol{B}_{\boldsymbol{z}_+} \mid \boldsymbol{B}_{\boldsymbol{x}_+} \otimes \boldsymbol{1}_{n_{\boldsymbol{z}_+}} \mid \boldsymbol{B}_{\boldsymbol{x}_+} \otimes \boldsymbol{B}_{\boldsymbol{z}_+}], \tag{4.6}$$

of dimension $n \times (1 + c_{\boldsymbol{z}_+} + c_{\boldsymbol{x}_+} + c_{\boldsymbol{z}_+} c_{\boldsymbol{x}_+})$, and where $\boldsymbol{1}_{n_{\boldsymbol{z}_+}}$ and $\boldsymbol{1}_{n_{\boldsymbol{x}_+}}$ are column vectors of ones of length $n_{\boldsymbol{z}_+}$ and $n_{\boldsymbol{x}_+}$ respectively. Therefore, model (4.5) is written as:

$$\boldsymbol{y}_+ = \gamma \boldsymbol{1}_{n_+} + (\boldsymbol{1}_{n_{\boldsymbol{x}_+}} \otimes \boldsymbol{B}_{\boldsymbol{z}_+}) \boldsymbol{\theta}_{\boldsymbol{z}_+} + (\boldsymbol{B}_{\boldsymbol{x}_+} \otimes \boldsymbol{1}_{n_{\boldsymbol{z}_+}}) \boldsymbol{\theta}_{\boldsymbol{x}_+} + (\boldsymbol{B}_{\boldsymbol{x}_+} \otimes \boldsymbol{B}_{\boldsymbol{z}_+}) \boldsymbol{\theta}_{\boldsymbol{s}_+},$$

with extended penalty the following block-diagonal matrix:

$$\boldsymbol{P}_+ = \begin{bmatrix} 0 & & & \\ & \lambda_{\boldsymbol{z}} \boldsymbol{D}'_{\boldsymbol{z}_+} \boldsymbol{D}_{\boldsymbol{z}_+} & & \\ & & \lambda_{\boldsymbol{x}} \boldsymbol{D}'_{\boldsymbol{x}_+} \boldsymbol{D}_{\boldsymbol{x}_+} & \\ & & & \tau_{\boldsymbol{z}} \boldsymbol{I}_{c_{\boldsymbol{x}_+}} \otimes \boldsymbol{D}'_{\boldsymbol{z}_+} \boldsymbol{D}_{\boldsymbol{z}_+} + \tau_{\boldsymbol{x}} \boldsymbol{D}'_{\boldsymbol{x}_+} \boldsymbol{D}_{\boldsymbol{x}_+} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}_+}} \end{bmatrix}, \tag{4.7}$$

of dimension $(1 + c_{\boldsymbol{z}_+} + c_{\boldsymbol{x}_+} + c_{\boldsymbol{z}_+} c_{\boldsymbol{x}_+}) \times (1 + c_{\boldsymbol{z}_+} + c_{\boldsymbol{x}_+} + c_{\boldsymbol{z}_+} c_{\boldsymbol{x}_+})$, where each block corresponds to the penalty over each of the coefficients of the model.

Although we have given expressions for extended basis and penalties, model (4.5) can not be fitted with (4.6) since it will yield singular matrices. Therefore, we will reformulate the extended P-spline S-ANOVA model (4.5) as a mixed model. To do so, we need to define an extended transformation matrix. As in Section 3.3, we can use the natural extended transformation matrix based on the SVD of the extended difference matrices, $\boldsymbol{\Omega}_+$, or an extended transformation matrix, $\boldsymbol{\Omega}^*_+$, that allow us to obtain extended mixed

model matrices that are direct extensions of the model matrices that give the fit.

The S-ANOVA model given in (4.1) is not identifiable but its mixed model representation allow us to impose easily the necessary constraints. To obtain predictions with S-ANOVA models subject to the constraint that the fit is maintained we have to use the extended transformation matrix $\boldsymbol{\Omega}_+^*$. In the following two sections we define the transformation matrices $\boldsymbol{\Omega}_+$ and $\boldsymbol{\Omega}_+^*$ and the model components associated to each case.

4.2.1 Natural reparametization of the S-ANOVA model into a mixed model for prediction

To reparameterize (4.5) as a mixed model the natural extended transformation matrix is $\boldsymbol{\Omega}_+ = [\boldsymbol{\Omega}_{+f} \mid \boldsymbol{\Omega}_{+r}]$ with dimension $(1 + c_{\boldsymbol{z}_+} + c_{\boldsymbol{x}_+} + c_{\boldsymbol{z}_+} c_{\boldsymbol{x}_+})$, where:

$$\boldsymbol{\Omega}_{+f} = \begin{bmatrix} 1 & & & \\ & 1 \otimes \boldsymbol{u}^{(2)}_{\boldsymbol{z}_+ f} & & \\ & & \boldsymbol{u}^{(2)}_{\boldsymbol{x}_+ f} \otimes 1 & \\ & & & \boldsymbol{u}^{(2)}_{\boldsymbol{x}_+ f} \otimes \boldsymbol{u}^{(2)}_{\boldsymbol{z}_+ f} \end{bmatrix},$$

$$\boldsymbol{\Omega}_{+r} = \begin{bmatrix} 1 & & & \\ & 1 \otimes \boldsymbol{U}_{\boldsymbol{z}_+ r} & & \\ & & \boldsymbol{U}_{\boldsymbol{x}_+ r} \otimes 1 & \\ & & & \boldsymbol{u}^{(2)}_{\boldsymbol{x}_+ f} \otimes \boldsymbol{U}_{\boldsymbol{z}_+ r} \mid \boldsymbol{U}_{\boldsymbol{x}_+ r} \otimes \boldsymbol{u}^{(2)}_{\boldsymbol{z}_+ f} \mid \boldsymbol{U}_{\boldsymbol{x}_+ r} \otimes \boldsymbol{U}_{\boldsymbol{z}_+ r} \end{bmatrix}, \tag{4.8}$$

where $\boldsymbol{u}^{(2)}_{\boldsymbol{z}_+ f}$ and $\boldsymbol{u}^{(2)}_{\boldsymbol{x}_+ f}$ are the second columns of $\boldsymbol{U}_{\boldsymbol{z}_+ f}$ and $\boldsymbol{U}_{\boldsymbol{x}_+ f}$, respectively, and $\boldsymbol{U}_{if}$ and $\boldsymbol{U}_{ir}$ are the eigenvectors corresponding to the zero values and positive values of the SVD of $\boldsymbol{D}_i' \boldsymbol{D}_i$, respectively, for $i = \boldsymbol{z}_+, \boldsymbol{x}_+$.

We obtain the fixed effects matrix as:

$$\boldsymbol{X}_+ = \boldsymbol{B}_+ \boldsymbol{\Omega}_{+f} = [\mathbf{1}_{n_+} \mid \mathbf{1}_{n_{\boldsymbol{x}_+}} \otimes \tilde{\boldsymbol{z}}_+ \mid \tilde{\boldsymbol{x}}_+ \otimes \mathbf{1}_{n_{\boldsymbol{z}_+}} \mid \tilde{\boldsymbol{x}}_+ \otimes \tilde{\boldsymbol{z}}_+], \tag{4.9}$$

where $\tilde{\boldsymbol{z}}_+$ and $\tilde{\boldsymbol{x}}_+$ are $\boldsymbol{B}_{\boldsymbol{z}_+} \boldsymbol{u}^{(2)}_{\boldsymbol{z}_+ f}$ and $\boldsymbol{B}_{\boldsymbol{x}_+} \boldsymbol{u}^{(2)}_{\boldsymbol{x}_+ f}$, respectively. The random effects matrix is obtained as:

$$\boldsymbol{Z}_+ = \boldsymbol{B}_+ \boldsymbol{\Omega}_{+_r} = [\mathbf{1}_{n_+} \mid \mathbf{1}_{n_{\boldsymbol{x}_+}} \otimes \boldsymbol{Z}_{\boldsymbol{z}_+} \mid \boldsymbol{Z}_{\boldsymbol{x}_+} \otimes \mathbf{1}_{n_{\boldsymbol{z}_+}} \mid \tilde{\boldsymbol{x}}_+ \otimes \boldsymbol{Z}_{\boldsymbol{z}_+} \mid \boldsymbol{Z}_{\boldsymbol{x}_+} \otimes \tilde{\boldsymbol{z}}_+ \mid \boldsymbol{Z}_{\boldsymbol{x}_+} \otimes \boldsymbol{Z}_{\boldsymbol{z}_+}], \tag{4.10}$$

where $\boldsymbol{Z}_{\boldsymbol{z}_+}$ and $\boldsymbol{Z}_{\boldsymbol{x}_+}$ are $\boldsymbol{B}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r}$ and $\boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}$, respectively.

For the extended transformation matrix given in (4.8) and extended penalty given in (4.7) the mixed model precision matrix of the random effects is given by the following theorem.

Theorem 4.1. *The extended precision matrix of random effects for S-ANOVA model in (4.5) with extended transformation matrix given in (4.8) and extended penalty given in (4.7) is the block-diagonal defined by:*

$$\boldsymbol{F}_+ = \text{blockdiag}(\boldsymbol{F}_+^{(1)}, \boldsymbol{F}_+^{(2)}, \boldsymbol{F}_+^{(1,2)}), \tag{4.11}$$

where for a second order penalty, it has size $(c_{\boldsymbol{z}_+}c_{\boldsymbol{x}_+} - 4) \times (c_{\boldsymbol{z}_+}c_{\boldsymbol{x}_+} - 4)$*, and where:*

$$\begin{aligned}
\boldsymbol{F}_+^{(1)} &= \lambda_{\boldsymbol{z}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}_+}, \\
\boldsymbol{F}_+^{(2)} &= \lambda_{\boldsymbol{x}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_+}, \\
\boldsymbol{F}_+^{(1,2)} &= \text{blockdiag}(\tau_{\boldsymbol{z}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}_+}, \tau_{\boldsymbol{x}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_+}, \tau_{\boldsymbol{z}}\boldsymbol{I}_{c_{\boldsymbol{z}_+}-2} \otimes \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}_+} + \tau_{\boldsymbol{x}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_+} \otimes \boldsymbol{I}_{c_{\boldsymbol{x}_+}-2}).
\end{aligned}$$

where $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}_+}$ *and* $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_+}$ *are the nonzero eigenvalues of the SVD of* $\boldsymbol{D}'_{\boldsymbol{z}_+}\boldsymbol{D}_{\boldsymbol{z}_+}$ *and* $\boldsymbol{D}'_{\boldsymbol{x}_+}\boldsymbol{D}_{\boldsymbol{x}_+}$*, respectively. The random effects covariance matrix is therefore* $\boldsymbol{G}_+ = \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{F}_+^{-1}$*.*

The proof of the previous theorem is given in Appendix C.1.

Once the model components are defined, the fit and the prediction are obtained simultaneously. The estimation of the fixed and random effects and the variance components would also be carried out by solving the extended Henderson system of equations (3.29) and maximizing the extended REML (3.30).

In the case in which just one covariate is extended, the models components are:

- Extended mixed model matrices:

$$\boldsymbol{X}_+ = \boldsymbol{B}_+\boldsymbol{\Omega}_{+f} = [\boldsymbol{1}_{n_+} \mid \boldsymbol{1}_{n_{\boldsymbol{x}_+}} \otimes \tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}}_+ \otimes \boldsymbol{1}_{n_{\boldsymbol{z}}} \mid \tilde{\boldsymbol{x}}_+ \otimes \tilde{\boldsymbol{z}}], \tag{4.12}$$

  where $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{x}}_+$ are $\boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{u}^{(2)}_{\boldsymbol{z}f}$ and $\boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{u}^{(2)}_{\boldsymbol{x}_+f}$, respectively. And

$$\boldsymbol{Z}_+ = \boldsymbol{B}_+\boldsymbol{\Omega}_{+r} = [\boldsymbol{1}_{n_+} \mid \boldsymbol{1}_{n_{\boldsymbol{x}_+}} \otimes \boldsymbol{Z}_{\boldsymbol{z}} \mid \boldsymbol{Z}_{\boldsymbol{x}_+} \otimes \boldsymbol{1}_{n_{\boldsymbol{z}}} \mid \tilde{\boldsymbol{x}}_+ \otimes \boldsymbol{Z}_{\boldsymbol{z}} \mid \boldsymbol{Z}_{\boldsymbol{x}_+} \otimes \tilde{\boldsymbol{z}} \mid \boldsymbol{Z}_{\boldsymbol{x}_+} \otimes \boldsymbol{Z}_{\boldsymbol{z}}] \tag{4.13}$$

  where $\boldsymbol{Z}_{\boldsymbol{z}}$ and $\boldsymbol{Z}_{\boldsymbol{x}_+}$ are $\boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}r}$ and $\boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}$, respectively.

- Extended random effects covariance matrix $\boldsymbol{G}_+ = \sigma^2_{\boldsymbol{\epsilon}} \boldsymbol{F}_+^{-1}$, with:

$$\boldsymbol{F}_+ = \text{blockdiag}(\boldsymbol{F}^{(1)}, \boldsymbol{F}_+^{(2)}, \boldsymbol{F}_+^{(1,2)}), \tag{4.14}$$

where for a second order penalty, $q_{\boldsymbol{x}} = q_{\boldsymbol{z}} = 2$, it has size $(c_{\boldsymbol{z}} c_{\boldsymbol{x}_+} - 4) \times (c_{\boldsymbol{z}} c_{\boldsymbol{x}_+} - 4)$, and where:

$$\begin{aligned}
\boldsymbol{F}^{(1)} &= \lambda_{\boldsymbol{z}} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}}, \\
\boldsymbol{F}_+^{(2)} &= \lambda_{\boldsymbol{x}} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_+}, \\
\boldsymbol{F}_+^{(1,2)} &= \text{blockdiag}(\tau_{\boldsymbol{z}} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}}, \tau_{\boldsymbol{x}} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_+}, \tau_{\boldsymbol{z}} \boldsymbol{I}_{c_{\boldsymbol{z}} - q_{\boldsymbol{z}}} \otimes \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} + \tau_{\boldsymbol{x}} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_+} \otimes \boldsymbol{I}_{c_{\boldsymbol{x}_+} - q_{\boldsymbol{x}}}),
\end{aligned}$$

with $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}_+}$ the nonzero eigenvalues of the SVD of $\boldsymbol{D}_{\boldsymbol{z}}' \boldsymbol{D}_{\boldsymbol{z}}$ and $\boldsymbol{D}_{\boldsymbol{x}_+}' \boldsymbol{D}_{\boldsymbol{x}_+}$, respectively.

However, as we have shown in the previous chapter, the method presented above does not ensure the invariance of the fit if out-of-sample prediction is performed.

4.2.2 Coherent prediction with S-ANOVA model

To predict with S-ANOVA models subject to the constraint that the fit has to be maintained an extended transformation matrix that preserves the model matrices has to be used. For the case in which the two covariates are extended, we define the following extended transformation matrix $\boldsymbol{\Omega}_+^* = [\boldsymbol{\Omega}_{+f}^* \mid \boldsymbol{\Omega}_{+r}^*]$ with dimension $(1 + c_{\boldsymbol{z}_+} + c_{\boldsymbol{x}_+} + c_{\boldsymbol{z}_+} c_{\boldsymbol{x}_+})$, and:

$$\boldsymbol{\Omega}_{+f}^* = \begin{bmatrix} 1 & & & \\ & 1 \otimes \boldsymbol{u}_{\boldsymbol{z}_+ f}^{*(2)} & & \\ & & \boldsymbol{u}_{\boldsymbol{x}_+ f}^{*(2)} \otimes 1 & \\ & & & \boldsymbol{u}_{\boldsymbol{x}_+ f}^{*(2)} \otimes \boldsymbol{u}_{\boldsymbol{z}_+ f}^{*(2)} \end{bmatrix},$$

$$\boldsymbol{\Omega}_{+r}^* = \begin{bmatrix} 1 & & & \\ & 1 \otimes \boldsymbol{U}_{\boldsymbol{z}_+ r}^* & & \\ & & \boldsymbol{U}_{\boldsymbol{x}_+ r}^* \otimes 1 & \\ & & & \boldsymbol{u}_{\boldsymbol{x}_+ f}^{*(2)} \otimes \boldsymbol{U}_{\boldsymbol{z}_+ r}^* \mid \boldsymbol{U}_{\boldsymbol{x}_+ r}^* \otimes \boldsymbol{u}_{\boldsymbol{z}_+ f}^{*(2)} \mid \boldsymbol{U}_{\boldsymbol{x}_+ r}^* \otimes \boldsymbol{U}_{\boldsymbol{z}_+ r}^* \end{bmatrix}, \tag{4.15}$$

where $\boldsymbol{u}_{\boldsymbol{z}_+ f}^{*(2)}$ and $\boldsymbol{u}_{\boldsymbol{x}_+ f}^{*(2)}$ are the second columns of $\boldsymbol{U}_{\boldsymbol{z}_+ f}^*$ and $\boldsymbol{U}_{\boldsymbol{x}_+ f}^*$, respectively, with $\boldsymbol{U}_{\boldsymbol{z}_+ f}^*$, $\boldsymbol{U}_{\boldsymbol{x}_+ f}^*$, $\boldsymbol{U}_{\boldsymbol{z}_+ r}^*$ and $\boldsymbol{U}_{\boldsymbol{x}_+ r}^*$ defined in (3.40) and (3.41).

The fixed effects matrix is:

$$\boldsymbol{X}_+^* = \boldsymbol{B}_+\boldsymbol{\Omega}_{+f}^* = [\mathbf{1}_{n_+} \mid \mathbf{1}_{n_{\boldsymbol{x}_+}} \otimes \tilde{\boldsymbol{z}}_+ \mid \tilde{\boldsymbol{x}}_+ \otimes \mathbf{1}_{n_{\boldsymbol{z}_+}} \mid \tilde{\boldsymbol{x}}_+ \otimes \tilde{\boldsymbol{z}}_+], \tag{4.16}$$

where $\tilde{\boldsymbol{z}}_+$ and $\tilde{\boldsymbol{x}}_+$ are $\boldsymbol{B}_{\boldsymbol{z}_+}\boldsymbol{u}_{\boldsymbol{z}f}^{*(2)}$ and $\boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{u}_{\boldsymbol{x}_+f}^{*(2)}$, respectively, and the random effects matrix is:

$$\boldsymbol{Z}_+^* = \boldsymbol{B}_+\boldsymbol{\Omega}_{+r}^* = [\mathbf{1}_{n_+} \mid \mathbf{1}_{n_{\boldsymbol{x}_+}} \otimes \boldsymbol{Z}_{\boldsymbol{z}_+}^* \mid \boldsymbol{Z}_{\boldsymbol{x}_+}^* \otimes \mathbf{1}_{n_{\boldsymbol{z}_+}} \mid \tilde{\boldsymbol{x}}_+ \otimes \boldsymbol{Z}_{\boldsymbol{z}_+}^* \mid \boldsymbol{Z}_{\boldsymbol{x}_+}^* \otimes \tilde{\boldsymbol{z}}_+ \mid \boldsymbol{Z}_{\boldsymbol{x}_+}^* \otimes \boldsymbol{Z}_{\boldsymbol{z}_+}^*], \tag{4.17}$$

where $\boldsymbol{Z}_{\boldsymbol{z}_+}^*$ and $\boldsymbol{Z}_{\boldsymbol{x}_+}^*$ are $\boldsymbol{B}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r}^*$ and $\boldsymbol{B}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}^*$, respectively. The covariance matrix of random effects is $\boldsymbol{G}_+^* = \sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{F}_+^{*-1}$, with $\boldsymbol{F}_+^*$ given by the following Theorem.

Theorem 4.2. *The extended precision matrix of random effects for S-ANOVA model in (4.5) with extended transformation matrix given in (4.15) and extended penalty given in (4.7) is the block-diagonal defined by:*

$$\boldsymbol{F}_+^* = \text{blockdiag}(\boldsymbol{F}^{(1)}, \boldsymbol{F}_+^{(2)}, \boldsymbol{F}_+^{(1,2)}), \tag{4.18}$$

where for a second order penalty, it has size $(c_{\boldsymbol{z}}c_{\boldsymbol{x}_+} - 4) \times (c_{\boldsymbol{z}}c_{\boldsymbol{x}_+} - 4)$*, and where:*

$$\begin{aligned}
\boldsymbol{F}^{(1)} &= \lambda_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{z}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r}^*, \\
\boldsymbol{F}_+^{(2)} &= \lambda_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{x}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}^*, \\
\boldsymbol{F}_+^{(1,2)} &= \begin{bmatrix} \boldsymbol{F}_{+11}^{(1,2)} & \boldsymbol{O} & \boldsymbol{F}_{+13}^{(1,2)} \\ \boldsymbol{O} & \boldsymbol{F}_{+22}^{(1,2)} & \boldsymbol{F}_{+23}^{(1,2)} \\ \boldsymbol{F}_{+13}^{(1,2)\prime} & \boldsymbol{F}_{+23}^{(1,2)\prime} & \boldsymbol{F}_{+33}^{(1,2)} \end{bmatrix},
\end{aligned}$$

with

$$\begin{aligned}
\boldsymbol{F}_{+11}^{(1,2)} &= \tau_{\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{x}_+f}^{*(2)\prime}\boldsymbol{u}_{\boldsymbol{x}_+f}^{*(2)} \otimes \boldsymbol{U}_{\boldsymbol{z}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{z}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r}^* \\
\boldsymbol{F}_{+13}^{(1,2)} &= \tau_{\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{x}_+f}^{*(2)\prime}\boldsymbol{U}_{\boldsymbol{x}_+r}^* \otimes \boldsymbol{U}_{\boldsymbol{z}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{z}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r}^*, \\
\boldsymbol{F}_{+22}^{(1,2)} &= \tau_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{x}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}^* \otimes \boldsymbol{u}_{\boldsymbol{z}_+r}^{*(2)\prime}\boldsymbol{u}_{\boldsymbol{z}_+r}^{*(2)}, \\
\boldsymbol{F}_{+23}^{(1,2)} &= \tau_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{x}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}^* \otimes \boldsymbol{u}_{\boldsymbol{z}_+f}^{*(2)\prime}\boldsymbol{U}_{\boldsymbol{z}_+r}^*. \\
\boldsymbol{F}_{+33}^{(1,2)} &= \tau_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\boldsymbol{U}_{\boldsymbol{x}_+r}^* \otimes \boldsymbol{U}_{\boldsymbol{z}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{z}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r}^* + \tau_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{x}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}^* \otimes \boldsymbol{U}_{\boldsymbol{z}_+r}^{*\prime}\boldsymbol{U}_{\boldsymbol{z}_+r}^*.
\end{aligned}$$

The proof is given in Appendix C.2.

If just one covariate is extended the extended mixed model components are:

- Extended mixed model matrices:

$$\boldsymbol{X}_{+}^{*} = [\mathbf{1}_{n_{+}} \mid \mathbf{1}_{n_{\boldsymbol{x}_{+}}} \otimes \tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{x}}_{+} \otimes \mathbf{1}_{n_{\boldsymbol{z}}} \mid \tilde{\boldsymbol{x}}_{+} \otimes \tilde{\boldsymbol{z}}], \tag{4.19}$$

where $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{x}}_{+}$ are $\boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{z}f}^{(2)}$ and $\boldsymbol{B}_{\boldsymbol{x}_{+}}\boldsymbol{u}_{\boldsymbol{x}_{+}f}^{*(2)}$, respectively. The random effects matrix is $\boldsymbol{Z}_{+}^{*} = \boldsymbol{B}_{+}\boldsymbol{\Omega}_{+r}^{*}$, i.e.:

$$\boldsymbol{Z}_{+}^{*} = [\mathbf{1}_{n_{+}} \mid \mathbf{1}_{n_{\boldsymbol{x}_{+}}} \otimes \boldsymbol{Z}_{\boldsymbol{z}} \mid \boldsymbol{Z}_{\boldsymbol{x}_{+}}^{*} \otimes \mathbf{1}_{n_{\boldsymbol{z}}} \mid \tilde{\boldsymbol{x}}_{+} \otimes \boldsymbol{Z}_{\boldsymbol{z}} \mid \boldsymbol{Z}_{\boldsymbol{x}_{+}}^{*} \otimes \tilde{\boldsymbol{z}} \mid \boldsymbol{Z}_{\boldsymbol{x}_{+}}^{*} \otimes \boldsymbol{Z}_{\boldsymbol{z}}], \tag{4.20}$$

where $\boldsymbol{Z}_{\boldsymbol{z}}$ and $\boldsymbol{Z}_{\boldsymbol{x}_{+}}$ are $\boldsymbol{B}_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}r}$ and $\boldsymbol{B}_{\boldsymbol{x}_{+}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*}$, respectively.

- Extended random effects covariance matrix $\boldsymbol{G}_{+}^{*} = \sigma_{\boldsymbol{\epsilon}}^{2}\boldsymbol{F}_{+}^{*-1}$ with

$$\boldsymbol{F}_{+}^{*} = \text{blockdiag}(\boldsymbol{F}^{(1)}, \boldsymbol{F}_{+}^{(2)}, \boldsymbol{F}_{+}^{(1,2)}), \tag{4.21}$$

where for a second order penalty, $q_{\boldsymbol{x}} = q_{\boldsymbol{z}} = 2$, it has size $(c_{\boldsymbol{z}}c_{\boldsymbol{x}_{+}} - 4) \times (c_{\boldsymbol{z}}c_{\boldsymbol{x}_{+}} - 4)$, and where:

$$\begin{aligned}
\boldsymbol{F}^{(1)} &= \lambda_{\boldsymbol{z}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}}, \\
\boldsymbol{F}_{+}^{(2)} &= \lambda_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*\prime}\boldsymbol{D}_{\boldsymbol{x}_{+}}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_{+}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*}, \\
\boldsymbol{F}_{+}^{(1,2)} &= \begin{bmatrix} \boldsymbol{F}_{+11}^{(1,2)} & \boldsymbol{O} & \boldsymbol{F}_{+13}^{(1,2)} \\ \boldsymbol{O} & \boldsymbol{F}_{+22}^{(1,2)} & \boldsymbol{O} \\ \boldsymbol{F}_{+13}^{(1,2)\prime} & \boldsymbol{O} & \boldsymbol{F}_{+33}^{(1,2)} \end{bmatrix},
\end{aligned}$$

with

$$\begin{aligned}
\boldsymbol{F}_{+11}^{(1,2)} &= \tau_{\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{x}_{+}f}^{*(2)\prime}\boldsymbol{u}_{\boldsymbol{x}_{+}f}^{*(2)} \otimes \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} \\
\boldsymbol{F}_{+13}^{(1,2)} &= \tau_{\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{x}_{+}f}^{*(2)\prime}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*} \otimes \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}}, \\
\boldsymbol{F}_{+22}^{(1,2)} &= \tau_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*\prime}\boldsymbol{D}_{\boldsymbol{x}_{+}}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_{+}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*} \otimes \boldsymbol{I}_{q_{\boldsymbol{z}}}, \\
\boldsymbol{F}_{+33}^{(1,2)} &= \tau_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*\prime}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*} \otimes \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} + \tau_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*\prime}\boldsymbol{D}_{\boldsymbol{x}_{+}}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_{+}}\boldsymbol{U}_{\boldsymbol{x}_{+}r}^{*} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}-q_{\boldsymbol{z}}}.
\end{aligned}$$

where $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ is the nonzero eigenvalues of the SVD of $\boldsymbol{D}_{\boldsymbol{z}}^{\prime}\boldsymbol{D}_{\boldsymbol{z}}$.

4.2.3 Prediction of aboveground biomass

In this section, we apply the proposed 2D interaction P-spline and S-ANOVA models to the real data set used in Section 2.3.1. We propose to predict out-of-sample aboveground biomass as a smooth function of height and diameter using the following two models:

- 2D interaction P-spline model:

$$\boldsymbol{y}_+ = f(\boldsymbol{z}_+, \boldsymbol{x}_+) + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+)$$

- Smooth-ANOVA model:

$$\boldsymbol{y}_+ = f(\boldsymbol{z}_+) + f(\boldsymbol{x}_+) + f(\boldsymbol{z}_+, \boldsymbol{x}_+) + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+)$$

where $\boldsymbol{y}_+$ is the extended response variable, the aboveground biomass, and $\boldsymbol{z}_+$ and $\boldsymbol{x}_+$ the extended covariates, the diameter and the height.

We have predicted weight for 10 new out-of-sample values for diameter and height. In Figure 4.1 we plot the smooth trend for height (left panel) and for diameter (right panel) obtained after fitting and predicting with the S-ANOVA model imposing that the fit is maintained. As it shows, both smooth terms are significantly different from zero. Figure 4.2 shows the fitted and predicted interaction function ($f(\boldsymbol{z}_+, \boldsymbol{x}_+)$) for the restricted S-ANOVA model.

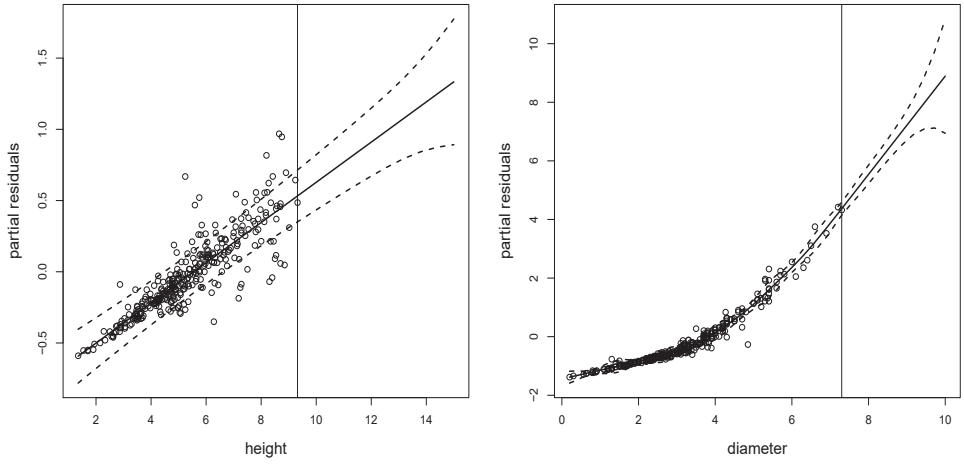


Figure 4.1: Fitted and predicted smooth curves for height (left panel) and for diameter (right panel) using the restricted S-ANOVA model. The vertical line indicates the height and diameter values from which we predict (9.32 and 7.3).

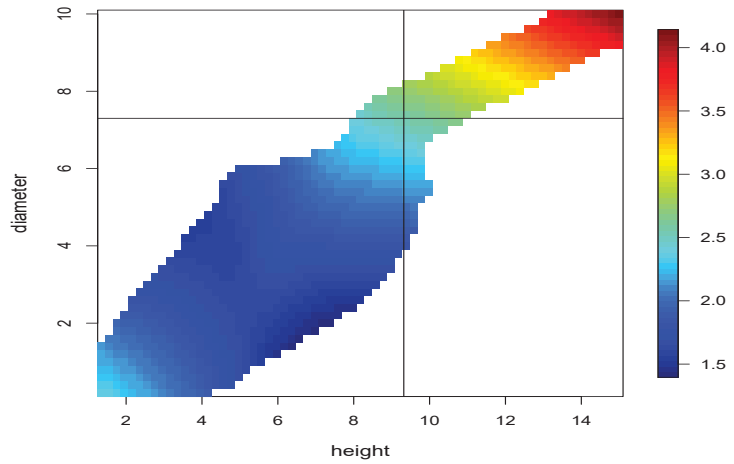


Figure 4.2: Fitted and predicted interaction function for the the restricted S-ANOVA model. The vertical line indicates the height value from which we predict (9.32) and the horizontal line indicates the diameter value from which we predict (7.3).

Figure 4.3 illustrates the prediction when the fit and the out-of-sample predictions are obtained with the S-ANOVA model (top left panel), with the S-ANOVA model imposing that the fit is maintained (top right panel), with the 2D interaction P-spline model (bottom left panel) and with the 2D interaction P-spline models imposing that the fit is maintained (bottom right panel). As the figure shows, the results obtained from the S-ANOVA model and from the restricted S-ANOVA model are almost equal. However, the solutions obtained from the 2D interaction P-spline models are different depending on if the restriction the fit is maintained is imposed or not. As we can appreciate, the fit changes significantly if the restriction is not imposed. In the prediction, the most significant difference is that the 2D interaction P-spline model gives lower weight values for the largest diameter and height values than the 2D interaction P-spline model. It is important to note that, in the fit, the degrees of freedom are 32 for the 2D interaction P-spline model and 12 for the S-ANOVA model, i.e. with the 2D interaction P-spline model we are adding unnecessary complexity.

Comparing the S-ANOVA models and the 2D interaction P-spline models, we conclude that the most coherent solution is given by the S-ANOVA models, since they give the largest weight values for the highest values of height and diameter. This conclusion will be reaffirmed in the next section, where we will see in a simulation study that, in most scenarios, the restricted S-ANOVA model outperforms the 2D interaction P-spline model in terms of accuracy.

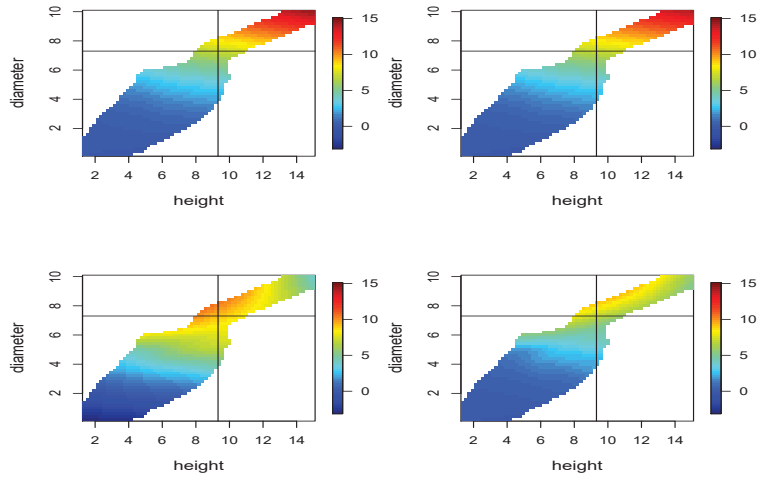


Figure 4.3: Fit and prediction with the S-ANOVA model (top left panel), with the S-ANOVA model imposing that the fit is maintained (top right panel), with the 2D interaction P-spline model (bottom left panel) and with the 2D interaction P-spline models imposing that the fit is maintained (bottom right panel) at out-of-sample values of diameter ($[7.3, 10]$) and height ($[9.32, 15]$).

4.3 Simulation study

In the previous chapter and previous sections, we have shown how to predict with interaction models (from P-spline and mixed models points of view) and with Smooth-ANOVA models, and how to impose restrictions. In this section, we examine the performance of the interaction and Smooth-ANOVA models in comparison to interaction and Smooth-ANOVA models in which we impose the constraint that the fit has to be the same as the fit we obtain when only fitting the data. For this propose we have simulated the data in two different scenarios:

a) Scenario 1. From an interaction model:

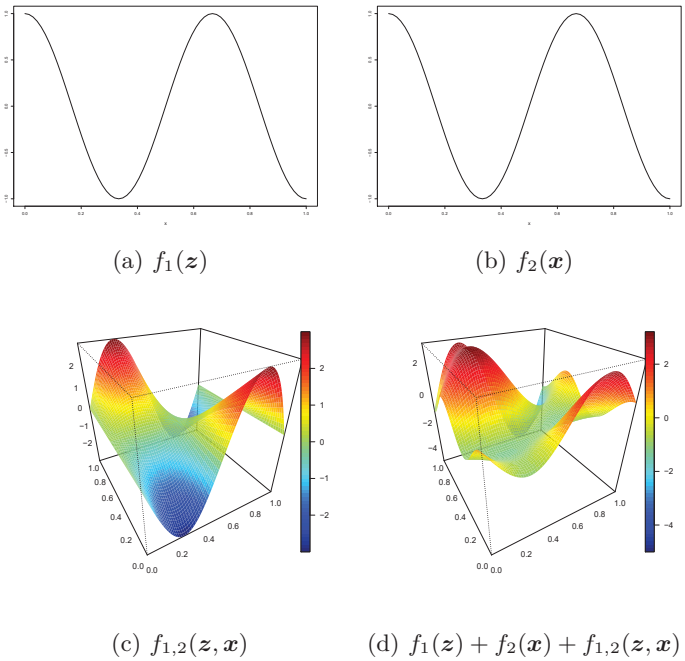
$$\boldsymbol{S}^1 = f_{1,2}(\boldsymbol{z}, \boldsymbol{x}).$$

b) Scenario 2. From a two main effects with interaction model:

$$\boldsymbol{S}^2 = f_1(\boldsymbol{z}) + f_2(\boldsymbol{x}) + f_{1,2}(\boldsymbol{z}, \boldsymbol{x}).$$

In both cases $f_1(\boldsymbol{z}) = \sin(2\pi\boldsymbol{z})$, $f_2(\boldsymbol{x}) = \cos(3\pi\boldsymbol{x})$ and $f_{1,2}(\boldsymbol{z}, \boldsymbol{x}) = 3\,\sin(2\pi\boldsymbol{z})(2\boldsymbol{x} - 1)$. To simulate the data we have generated a grid of 4900 values. Both covariates, $\boldsymbol{z}$ and $\boldsymbol{x}$, take 70 equidistant values in the interval $[0, 1]$, and the errors are independent and

identically distributed, with mean 0 and variance $\sigma^2_{\epsilon} = 0.25$. Figure 4.4 shows $f_1(\boldsymbol{z})$, $f_2(\boldsymbol{x})$ and the surfaces proposed in scenarios 1 and 2.



(a) $f_1(\boldsymbol{z})$ (b) $f_2(\boldsymbol{x})$

(c) $f_{1,2}(\boldsymbol{z}, \boldsymbol{x})$ (d) $f_1(\boldsymbol{z}) + f_2(\boldsymbol{x}) + f_{1,2}(\boldsymbol{z}, \boldsymbol{x})$

Figure 4.4: Functions (a) and (b) are the nonlinear main effects of $\boldsymbol{z}$ and $\boldsymbol{x}$, (c) is the interaction surface and (d) is the surface generated from the two main effects and the interaction surface.

For each scenario, we fit and predict with four models, the models and their components are listed below:

- 2D interaction P-spline model, i.e.

$$\begin{aligned} \boldsymbol{y}_+ &= f(\boldsymbol{z}_+, \boldsymbol{x}_+) + \boldsymbol{\epsilon}_+ \\ &= \boldsymbol{X}_+ \boldsymbol{\beta}_{\tilde{+}} + \boldsymbol{Z}_+ \boldsymbol{\alpha}_+ + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+), \quad \boldsymbol{\alpha}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}_+), \end{aligned}$$

  where the model components depend on how many covariates are extended:

  - Extending one covariate, the model components are $\boldsymbol{X}_+$ and $\boldsymbol{Z}_+$ defined in (3.36) and (3.37), $\boldsymbol{R}_+$ defined in (3.15) and $\boldsymbol{G}_+$ defined in (3.38).
  - Extending two covariates, the model components are $\boldsymbol{X}_+$ and $\boldsymbol{Z}_+$ defined in (3.32) and (3.33), $\boldsymbol{R}_+$ defined in (3.10) and $\boldsymbol{G}_+$ defined in (3.34).

  In both cases, after defining the model components, to estimate the model we maximize the extended REML (3.30) and solve the extended mixed model system of

equations of Henderson (3.29) to estimate the model parameters. This can be done through the SOP algorithm implemented by Rodríguez-Álvarez et al. (2018), since we can give infinite variance to the unknown values, i.e. we can use $\tilde{\boldsymbol{R}}_+^{-1}$ as the inverse of the variance-covariance matrix of the error through a matrix of weights.

- 2D interaction P-spline model with restriction, i.e.

$$\begin{aligned} \boldsymbol{y}_+ &= f(\boldsymbol{z}_+, \boldsymbol{x}_+) + \boldsymbol{\epsilon}_+ \\ &= \boldsymbol{X}_+\boldsymbol{\beta}_+^* + \boldsymbol{Z}_+\boldsymbol{\alpha}_+^* + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+), \quad \boldsymbol{\alpha}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}_+), \end{aligned}$$

subject to the fit has to be maintained, restriction imposed through a equation $\boldsymbol{C}_{MM}\begin{bmatrix}\boldsymbol{\beta}_+^* \\ \boldsymbol{\alpha}_+^*\end{bmatrix} = \boldsymbol{r}_{MM}$, where the model components depend on how many covariates are extended:

  - Extending one covariate, the model components are $\boldsymbol{X}_+$ and $\boldsymbol{Z}_+$ defined in (3.36) and (3.37), $\boldsymbol{R}_+$ defined in (3.15) and $\boldsymbol{G}_+$ defined in (3.38).
  - Extending two covariates, the model components are $\boldsymbol{X}_+$ and $\boldsymbol{Z}_+$ defined in (3.32) and (3.33), $\boldsymbol{R}_+$ defined in (3.10) and $\boldsymbol{G}_+$ defined in (3.34).

In both cases, we define the constraints matrix $\boldsymbol{C}_{MM}$ and the constraints vector $\boldsymbol{r}_{MM}$ in the P-splines context and use the extended transformations ($\boldsymbol{\Omega}_+$ given in (3.35) if we extend one covariate and $\boldsymbol{\Omega}_+$ given in (3.31) if we extend the two covariates) to obtain the constraints matrix in the context of mixed models, $\boldsymbol{C}_{MM} = \boldsymbol{C}\boldsymbol{\Omega}_+$. Since we are imposing the restriction that the fit has to be maintained, the covariance parameters used to obtain the fit and the prediction simultaneously are the ones estimated to compute the fit. Once the model components are defined the fixed and random effects are computed solving the system (3.47).

Notice that if we impose invariance of the fit the variance components used are the ones from the fit, while if we do not impose that restriction, the variance components are the ones that maximize the extended REML. From our experience, it is worth saying that the solutions of the REML and of the extended REML are not very different.

- Smooth-ANOVA model, i.e.

$$\begin{aligned} \boldsymbol{y}_+ &= f_1(\boldsymbol{z}_+) + f_2(\boldsymbol{x}_+) + f_{1,2}(\boldsymbol{z}_+, \boldsymbol{x}_+) + \boldsymbol{\epsilon}_+ \\ &= \boldsymbol{X}_+\boldsymbol{\beta}_+^{\tilde{}} + \boldsymbol{Z}_+\boldsymbol{\alpha}_+ + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+), \quad \boldsymbol{\alpha}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}_+), \end{aligned}$$

  where the model components are defined depending on how many covariates we extend.

  - Extending one covariate, the components are $\boldsymbol{X}_+$ and $\boldsymbol{Z}_+$ defined in (4.12) and (4.13), $\boldsymbol{R}_+$ defined in (3.15) and $\boldsymbol{G}_+$ defined through the equation of $\boldsymbol{F}_+$ in (4.14).
  - Extending two covariates, the components are $\boldsymbol{X}_+$ and $\boldsymbol{Z}_+$ defined in (4.9) and (4.10), $\boldsymbol{R}_+$ defined in (3.10) and $\boldsymbol{G}_+$ defined through the equation of $\boldsymbol{F}_+$ in (4.11).

  In both cases, after defining the model components, to estimate the model we maximize the extended REML (3.30) and solve the extended mixed model system of equations of Henderson (3.29).

- Smooth-ANOVA model with restriction, i.e.

$$\begin{aligned} \boldsymbol{y}_+ &= f_1(\boldsymbol{z}_+) + f_2(\boldsymbol{x}_+) + f_{1,2}(\boldsymbol{z}_+, \boldsymbol{x}_+) + \boldsymbol{\epsilon}_+ \\ &= \boldsymbol{X}_+\boldsymbol{\beta}_+^{*} + \boldsymbol{Z}_+\boldsymbol{\alpha}_+^{*} + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_+), \quad \boldsymbol{\alpha}_+^{*} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}_+), \end{aligned}$$

  subject to the fit has to be maintained, restriction imposed through a equation $\boldsymbol{C}_{MM}\begin{bmatrix}\boldsymbol{\beta}_+^{*} \\ \boldsymbol{\alpha}_+^{*}\end{bmatrix} = \boldsymbol{r}_{MM}$, where the model components depend on how many covariates are extended:

  - Extending one covariate, the components are $\boldsymbol{X}_+$ and $\boldsymbol{Z}_+$ defined in (4.19) and (4.20), $\boldsymbol{R}_+$ defined in (3.15) and $\boldsymbol{G}_+$ defined through $\boldsymbol{F}_+$ in (4.21).
  - Extending two covariates, the components are $\boldsymbol{X}_+$ and $\boldsymbol{Z}_+$ defined in (4.16) and (4.17), $\boldsymbol{R}_+$ defined in (3.10) and $\boldsymbol{G}_+$ defined through $\boldsymbol{F}_+$ in (4.18).

  Again the variance parameters are the ones estimated to compute the fit. Once the model components are defined the fixed and random effects are computed solving the system (3.47).

For each scenario and each model, we have repeated the following 100 times:

- Start with a data set of observations arranged into a matrix $\boldsymbol{Y}_+$ of dimension $n_{\boldsymbol{z}_+} \times n_{\boldsymbol{x}_+}$:

$$\boldsymbol{Y}_+ = \begin{bmatrix} \boldsymbol{Y} & \boldsymbol{Y}_{\boldsymbol{z}\boldsymbol{x}_p} \\ \boldsymbol{Y}_{\boldsymbol{z}_p\boldsymbol{x}} & \boldsymbol{Y}_{\boldsymbol{z}_p\boldsymbol{x}_p} \end{bmatrix},$$

  where $\boldsymbol{Y}$, $\boldsymbol{Y}_{\boldsymbol{z}\boldsymbol{x}_p}$, $\boldsymbol{Y}_{\boldsymbol{z}_p\boldsymbol{x}}$ and $\boldsymbol{Y}_{\boldsymbol{z}_p\boldsymbol{x}_p}$ have dimension $n_{\boldsymbol{z}} \times n_{\boldsymbol{x}}$, $n_{\boldsymbol{z}} \times n_{\boldsymbol{x}_p}$, $n_{\boldsymbol{z}_p} \times n_{\boldsymbol{x}}$ and $n_{\boldsymbol{z}_p} \times n_{\boldsymbol{x}_p}$, respectively.

- Split our data into two groups: the training data $\boldsymbol{Y}$ and the test data $\boldsymbol{Y}_{\boldsymbol{z}\boldsymbol{x}_p}$, $\boldsymbol{Y}_{\boldsymbol{z}_p\boldsymbol{x}}$ and $\boldsymbol{Y}_{\boldsymbol{z}_p\boldsymbol{x}_p}$.

- Use the training data to predict $4900 - n$ new observations ($n = n_{\boldsymbol{z}} \times n_{\boldsymbol{x}}$).

- Check the accuracy of each model.

The marginal B-spline bases $\boldsymbol{B}_{\boldsymbol{z}}$ and $\boldsymbol{B}_{\boldsymbol{x}}$ are constructed with 15 knots and cubic splines and then extended to cover the whole range of $\boldsymbol{z}_+$ and $\boldsymbol{x}_+$, the penalty orders are two.

As it is stated in the literature (Hyndman et al. 2008) a model which fits the data well does not necessarily predict well, therefore we have compared the fit performance, the prediction performance and the overall performance of the methods. To check the accuracy of the methods we follow Hyndman (2006) and take the errors as the difference between the function values from which we simulate the data and the fit and prediction produced using only the data in the training set:

$$\boldsymbol{E}_+ = \boldsymbol{S}^k - \hat{\boldsymbol{Y}}_+, \quad k = 1, 2$$

i.e, we have the errors matrix $\boldsymbol{E}_+ = \begin{bmatrix} \boldsymbol{E} & \boldsymbol{E}_{\boldsymbol{z}\boldsymbol{x}_p} \\ \boldsymbol{E}_{\boldsymbol{z}_p\boldsymbol{x}} & \boldsymbol{E}_{\boldsymbol{z}_p\boldsymbol{x}_p} \end{bmatrix}$, and therefore the vectors containing the errors in the fit, in the prediction, and in the overall performance:

- Vector with the fit errors: $\boldsymbol{e}^{(f)} = \text{vec}(\boldsymbol{E})$
- Vector with the prediction errors: $\boldsymbol{e}^{(p)} = (\text{vec}(\boldsymbol{E}_{\boldsymbol{z}\boldsymbol{x}_p}), \text{vec}(\boldsymbol{E}_{\boldsymbol{z}_p\boldsymbol{x}}), \text{vec}(\boldsymbol{E}_{\boldsymbol{z}_p\boldsymbol{x}_p}))$
- Vector with the total errors: $\boldsymbol{e}^{(t)} = \text{vec}(\boldsymbol{E}_+)$

The errors measure that we use is the mean absolute error because as it is said in Hyndman (2006) it is less sensitive to outliers than the root mean square error:

Mean absolute error: $\text{MAE} = \frac{\sum_{i=1}^{N} |e_i^{(l)}|}{N}, \quad N = \text{length}(\boldsymbol{e}^{(l)})$

for $l = f, p, t$, i.e. for the errors in the fit, in the prediction or in total.

We have divided the results of the simulation study in two sections, one to show the obtained results in Scenario 1 and other to show the obtained results in Scenario 2.

4.3.1 Simulations results for Scenario 1

Below we show the obtained results for Scenario 1. For different values of $n_{\boldsymbol{z}_p}$ and $n_{\boldsymbol{x}_p}$, we fit and predict with the four smooth mixed models: interaction, interaction with restriction, S-ANOVA and S-ANOVA with restriction. We have illustrated the results with boxplots for MAE values of the fit, of the prediction and in the overall performance.

The x-axis labels of the boxpots refer to the models: *unrestricted* (2D interaction P-spline model), *restricted* (2D interaction P-spline model with restriction), *ANOVA* (S-ANOVA model) and *restricted ANOVA* (S-ANOVA model with restriction).

Notice that for the different values of $n_{\boldsymbol{z}_p}$ and $n_{\boldsymbol{x}_p}$ we consider as the training data an observations matrix of dimension $(70 - n_{\boldsymbol{z}_p}) \times (70 - n_{\boldsymbol{x}_p})$ and we predict $(70 - n_{\boldsymbol{z}_p}) \times n_{\boldsymbol{x}_p} + n_{\boldsymbol{z}_p} \times (70 - n_{\boldsymbol{x}_p}) + n_{\boldsymbol{z}_p} \times n_{\boldsymbol{x}_p}$ new values. The results obtained for this scenario depend on the dimensions in which we are predicting:

i) In the cases in which just one covariate is extended: in the fit, the performances of the restricted S-ANOVA model and of the 2D P-splines models (with and without imposing invariance in the fit) are very similar and these are the most accurate models (lower MAE values). In the prediction, the three models behave similarly, however, the performance of the restricted S-ANOVA model is a bit worse and its MAE error values increase with an increasing of the prediction horizon.

ii) In the cases in which the two covariates are extended: in the fit and in the prediction, the most accurate model is the restricted S-ANOVA model, moreover the 2D P-splines models are almost so accurate as it.

Moreover, the S-ANOVA model has degrees of freedom similar to or less than the 2D P-splines models, i.e. S-ANOVA models do not incorporate unnecessary complexity. All results are listed in Appendix C.3 for the different values of $n_{\boldsymbol{z}_p}$ and $n_{\boldsymbol{x}_p}$. To illustrate the previous conclusions, in Figure 4.5, we show the boxplots for the particular case of $n_{\boldsymbol{z}_p} = 10$ and $n_{\boldsymbol{x}_p} = 10$.

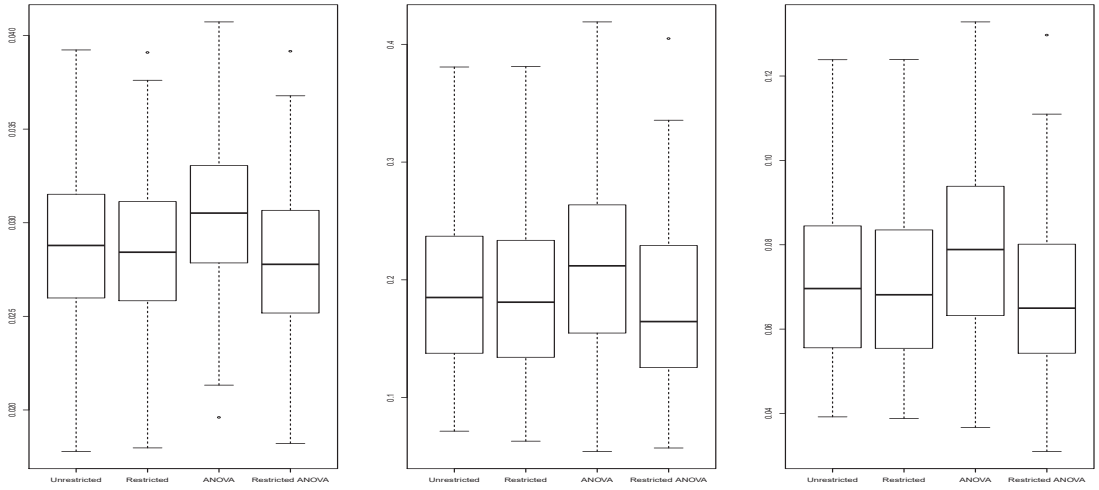


Figure 4.5: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 1 and $n_{\boldsymbol{z}_p} = 10$, $n_{\boldsymbol{x}_p} = 10$.

4.3.2 Simulations results for Scenario 2

In Appendix C.3 we show the results obtained for Scenario 2, i.e., for the case in which the true surface is constructed from a model with two main effects and with an interaction.

In this scenario, there is a significant difference between the results obtained with the interaction models and with the S-ANOVA models: 2D interaction P-splines models (restricted and unrestricted) are always less accurate than S-ANOVA models. This fact is due to interaction models are constraint to fit the true model without taking into account the main effects.

We can summarize the performances of the 2D P-spline interaction models (with and without restriction) as follows: in the fit both models are quite similar, however, in the prediction, the restricted interaction model is better than the unrestricted one, the major difference between the two models can be seen for most scenarios.

In the case of the S-ANOVA models, the restricted model is always better than the unrestricted one, in the fit, in the prediction, and therefore in the overall performance.

The conclusion for this scenario is that the restricted S-ANOVA model is clearly the most accurate. To illustrate this, Figure 4.6 shows the boxplots for the particular case of $n_{\boldsymbol{z}_p} = 20$ and $n_{\boldsymbol{x}_p} = 5$.

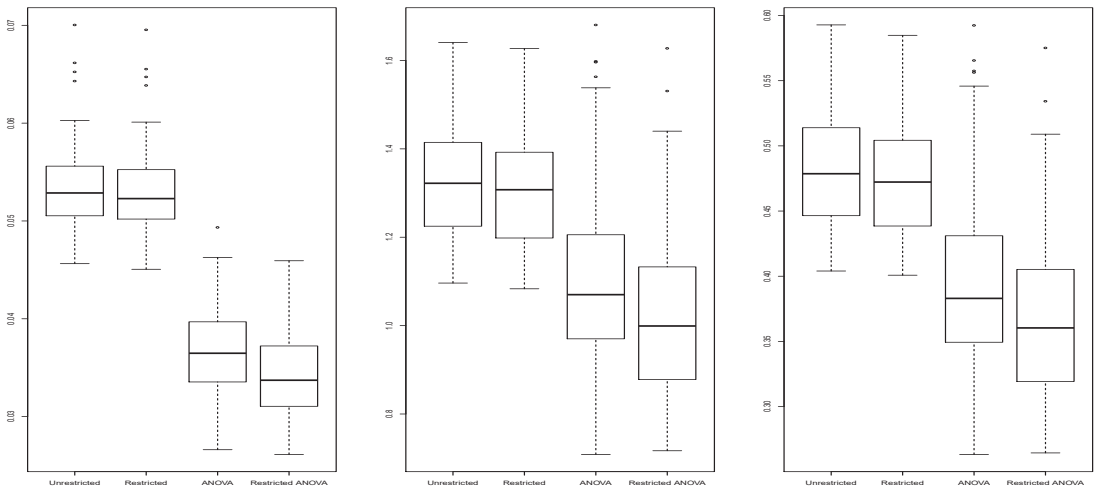


Figure 4.6: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 2 and $n_{\boldsymbol{z}_p} = 20$ and $n_{\boldsymbol{x}_p} = 5$.

Therefore, taking into account the results obtained for both scenarios, our suggestion in models with interaction is to use the restricted S-ANOVA model always.

4.4 Conclusions of the chapter

In this chapter we have developed a method for out-of-sample prediction for the Smooth-ANOVA model proposed in Lee and Durbán (2011). We have reviewed the methodology for fitting a smooth interaction function as a decomposition of smooth functions which are identifiable, and also defined the extended B-spline basis and penalty matrix to fit and predict simultaneously with a S-ANOVA model. In order to get an identifiable model, we have reparameterized the extended S-ANOVA model as a mixed model. The transformations used to rewrite a S-ANOVA model as a mixed model are not orthogonal, since with the transformation we drop the unnecessary terms, and hence, to impose restrictions over the coefficients we have to work in the mixed model context. Therefore, to impose invariance of the fit, it is crucial to define an extended transformation matrix that preserves the model matrices from the fit, we have done it in Section 4.2.2.

A simulation study has been carried out to compare constrained and unconstrained out-of-sample prediction when using 2D interaction models and S-ANOVA models. From the results of the simulation study, we have concluded that in most situations the constrained S-ANOVA model behaves better in the fit and out-of-sample predictions when the prediction is carried out in both dimensions of the interaction term. When prediction

is needed in only one of the covariates results depend on the simulated scenario, although in most cases the S-ANOVA model outperformed the full interaction (restricted or not) model.

The constrained prediction method proposed has also been used in one real data example to predict tree biomass as a function of the height and diameter of the tree.

Chapter 5

Prediction in penalized generalized linear models

The methodology described in the previous chapters can be extended to the case of data within the exponential family of distributions, i.e., within the generalized linear model (GLM) framework. GLMs are a class of statistical models that are a natural generalization of classical linear models to more general responses (such as Poisson or Binomial distributions). GLMs have a common method for the estimation of parameters by maximum likelihood; this uses weighted least squares with an adjusted dependent variate, and does not require preliminary guesses to be made of the parameter values. All the theory related to GLMs can be seen in McCullagh and Nelder (1989).

This chapter introduces the approach of prediction in Penalized Generalized Linear Models (P-GLMs). Section 5.1 is dedicated to briefly review the main characteristics of P-GLMs. We also introduce our proposal to predict out-of-sample values in P-GLMs. The methodology in the context of Penalized Generalized Linear Mixed Models (P-GLMMs) is shown in Section 5.2. In Section 5.3 a method to obtain predictions subject to a set of restrictions in the P-GLMs and the P-GLMMs framework is shown. Finally, in Section 5.4 an example illustrates the methodology with the analysis of US data on the number of deaths from respiratory disease.

5.1 Prediction in Penalized Generalized Linear Models

An important concept that unifies all GLMs is the exponential family of distributions. Let $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n)'$ be a vector of n observations whose components are independently distributed with means $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_n)'$, and each of them has a distribution in the exponential family, taking the form:

$$f(\boldsymbol{y}; \boldsymbol{\gamma}, \phi) = \exp\left\{ \frac{\boldsymbol{y}'\boldsymbol{\gamma} - \mathbf{1}_n' b(\boldsymbol{\gamma})}{a(\phi)} + \mathbf{1}_n' c\left(\boldsymbol{y}, \phi\right) \right\},$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_n)'$ are the canonical parameters, ϕ the dispersion canonical parameter, $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions and $\mathbf{1}_n$ denotes a vector of ones of length n. It can be shown that $\mathbb{E}[\boldsymbol{y}] = b'(\boldsymbol{\gamma})$ and $\text{Var}[\boldsymbol{y}] = \phi \cdot b''(\boldsymbol{\gamma})$, where $b(\boldsymbol{\gamma})$ and $b''(\boldsymbol{\gamma})$ are the first and second derivatives of b. The basic structure of a GLM is:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}), \text{ and } \boldsymbol{\mu} = \mathbb{E}[\boldsymbol{y}] = g^{-1}(\boldsymbol{\eta}), \tag{5.1}$$

where $\boldsymbol{\eta}$ is the linear predictor, and g is a monotonic differentiable function which relates the mean with the linear predictor, called link function. There are many choices of link functions and usually the canonical link is selected (i.e. a function g such that $\boldsymbol{\eta} = \boldsymbol{\gamma}$).

To fit a GLM in the framework of P-splines, we will assume that the linear predictor of the covariate $\boldsymbol{\eta} = f(\boldsymbol{x})$ may have a basis representation, in particular $\boldsymbol{\eta} = \boldsymbol{B\theta}$, with $\boldsymbol{B}$ a B-splines basis and $\boldsymbol{\theta}$ the coefficients vector (for more details see Eilers and Marx 1996 and Marx and Eilers 1999). In this context and under the canonical link ($\boldsymbol{\eta} = \boldsymbol{\gamma}$), the joint density is:

$$f(\boldsymbol{\theta}; \boldsymbol{y}) = \exp\left\{ \frac{\boldsymbol{y}'(\boldsymbol{B\theta}) - \mathbf{1}_n' b(\boldsymbol{B\theta})}{a(\phi)} + \mathbf{1}_n' c(\boldsymbol{y}, \phi) \right\}$$

Therefore, the penalized log-likelihood function is:

$$\mathcal{L}_p(\boldsymbol{\theta}; \boldsymbol{y}) = \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) - \tfrac{1}{2}\lambda \boldsymbol{\theta}' \boldsymbol{P\theta}, \tag{5.2}$$

where $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{\boldsymbol{y}'\boldsymbol{B\theta} - \mathbf{1}_n' b(\boldsymbol{B\theta})}{a(\phi)} + \mathbf{1}' c(\boldsymbol{y})$ is the ordinary log-likelihood function, and $\boldsymbol{P}$ is the penalty matrix. For simplicity, we ignore the role of the dispersion in the previous distribution and set parameter $a(\phi) = 1$. Differentiating (5.2) w.r.t. $\boldsymbol{\theta}$ we obtain the system of equations:

$$\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{\mu}) = \lambda \boldsymbol{P\theta}.$$

These equations are nonlinear in $\boldsymbol{\theta}$ and an iterative procedure is needed to solve them as, for example, penalized iteratively re-weighted least square (PIRLS) algorithm based on the Newton-Raphson method (for more details, see McCullagh and Nelder 1989). For a given penalty matrix $\lambda \boldsymbol{P}$, the penalized version of the scoring algorithm is:

$$(\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B} + \lambda \boldsymbol{P})\hat{\boldsymbol{\theta}} = \boldsymbol{B}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}, \tag{5.3}$$

where the matrix $\boldsymbol{W}$ is diagonal with elements $w_i = \{g'(\mu_i)\}^{-1}$ (in general $w_i = \{v(\mu_i)(g'(\mu_i))^2\}^{-1}$, with $v(\cdot)$ the variance function of the exponential family distribution given, but for the canonical link $g'(\mu_i) = v^{-1}(\mu_i)$, see McCullagh and Nelder 1989),

and $z_i = \eta_i + (y_i - \mu_i)g'(\mu_i)$ known as the working vector. We use the symbols ($\tilde{}$) and ($\hat{}$) to remark that are the update estimate and the current approximate solution, respectively.

The estimating PIRLS algorithm is summarized as follows:

---

1. set an initial value of $\boldsymbol{\theta}$ $(= \hat{\boldsymbol{\theta}}_{old})$
2. use $\hat{\boldsymbol{\theta}}_{old}$ to estimate $\boldsymbol{W}$ and $\boldsymbol{\mu}_{old}$
3. let $\hat{\boldsymbol{\eta}}_{old} = \boldsymbol{B}\hat{\boldsymbol{\theta}}_{old}$, get the $\boldsymbol{z}_{new}$
4. obtain the new estimate $\hat{\boldsymbol{\theta}}_{new}$
5. repeat 2 to 4 until the convergence criterion is satisfied $\left(\text{e.g. } \frac{\sum_{i=1}^{n}(\eta_{new_i} - \eta_{old_i})^2}{\sum_{i=1}^{n}\eta_{new_i}^2} < \delta\right)$ for a small value of δ (e.g. $\delta \simeq 10^{-6}$).

Note that $\boldsymbol{z}$ is just a linearized form of the link function applied to the data, to first order,

$$g(\boldsymbol{y}) \simeq g(\boldsymbol{\mu}) + (\boldsymbol{y} - \boldsymbol{\mu})g'(\boldsymbol{\mu}) = \boldsymbol{z}.$$

The variance of $\boldsymbol{z}$ is just $\boldsymbol{W}^{-1}$, assuming that $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ are fixed and known. Therefore, taking into account that in the previous estimation procedure the linearized variable and the diagonal matrix $\boldsymbol{W}$ change at each iteration, the procedure is analogous to the estimation in the associated normal theory model

$$\boldsymbol{z} = \boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{W}^{-1}).$$

As we will see in the next section, taking into account the previous associated theory model, the extension of the prediction methodology to the GLMs framework is straightforward.

5.1.1 Out-of-sample prediction

The methodology associated to a P-GLM can be used to predict new out-of-sample values. Suppose that given a vector $\boldsymbol{y}$ of n observations of the response variable we want to predict n_p new values $\boldsymbol{y}_p$ at $\boldsymbol{x_p}$. We define the extended linear predictor as $\boldsymbol{\eta}_+ = \boldsymbol{B}_+\boldsymbol{\theta}_+$, with extended basis $\boldsymbol{B}_+$ and extended vector of coefficients $\boldsymbol{\theta}_+$. Therefore,

the extended penalized log-likelihood function is:

$$\mathcal{L}_{p_+}(\boldsymbol{\theta}_+; \boldsymbol{y}_+) = \boldsymbol{y}'_+\boldsymbol{B}_+\boldsymbol{\theta}_+ - \mathbf{1}'_{n_+} b(\boldsymbol{B}_+\boldsymbol{\theta}_+) + \mathbf{1}'_+ c(\boldsymbol{y}_+) - \tfrac{1}{2}\lambda\boldsymbol{\theta}'_{n_+}\boldsymbol{P}_+\boldsymbol{\theta}_+, \qquad (5.4)$$

where $\boldsymbol{P}_+$ is the extended penalty matrix and $\boldsymbol{y}_+ = (\boldsymbol{y}', \boldsymbol{y}'_p)'$ with $\boldsymbol{y}_p$ arbitrary values. Differentiating (5.4) w.r.t. $\boldsymbol{\theta}_+$ we obtain the system of equations:

$$\boldsymbol{B}'_+(\boldsymbol{y}_+ - \boldsymbol{\mu}_+) = \lambda\boldsymbol{P}_+\boldsymbol{\theta}_+,$$

which, as for the fit, is nonlinear in $\boldsymbol{\theta}_+$.
Extending the idea that we have used in the previous chapters, since the $\boldsymbol{y}_p$ values of $\boldsymbol{y}_+$ are unknown, the natural way of adapting the idea of infinite variance to express that we do not have any information about the data $\boldsymbol{y}_p$, is to consider infinite variance for the values $\boldsymbol{z}_p$ of the extended linearized variable. That is, we consider $\boldsymbol{z}_+ = (\boldsymbol{z}', \boldsymbol{z}'_p)'$, with $\boldsymbol{z} = \boldsymbol{\eta} + (\boldsymbol{y} - \boldsymbol{\mu})g'(\boldsymbol{\mu})$ and $\boldsymbol{z}_p$ arbitrary values with infinite variance, i.e. $w_i^{-1} = \infty$ for $i = n+1, ..., n_+$ (notice that $w_i = 0$ for $i = n+1, ..., n_+$).

This yields the extended penalized version of the scoring algorithm:

$$(\boldsymbol{B}'_+\tilde{\boldsymbol{W}}_+\boldsymbol{B}_+ + \lambda\boldsymbol{P}_+)\hat{\boldsymbol{\theta}}_+ = \boldsymbol{B}'_+\tilde{\boldsymbol{W}}_+\tilde{\boldsymbol{z}}_+ \qquad (5.5)$$

where, $\tilde{\boldsymbol{z}}_+ = \begin{bmatrix} \boldsymbol{\eta} + (\boldsymbol{y} - \boldsymbol{\mu})g'(\boldsymbol{\mu}) \\ \boldsymbol{z}_p \end{bmatrix}$, with $\boldsymbol{z}_p$ arbitrary values, and $\boldsymbol{W}_+$ a diagonal matrix with elements $w_i = \{g'(\mu_i)\}^{-1}$ for $i = 1, ..., n$ and $w_i = 0$ for $i = n+1, ..., n_+$.
Following the proof given for the univariate Gaussian model (see Section 2.1.1), it is straightforward to prove that for a fixed smoothing parameter λ and $\boldsymbol{P}_+ = \boldsymbol{D}'_+\boldsymbol{D}_+$, the fit obtained through the extended scoring algorithm is the same as the solution we obtain only fitting the data.

5.2 Mixed models representation of P-GLM for prediction

As we have done for the approach under the normality case. If we consider P-GLMs, they can be represented as generalized linear mixed models (GLMMs) (Breslow and Clayton 1993). In a GLMM the linear predictor is defined as:

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha},$$

where $g(\cdot)$ is a link function defined as in (5.1) and $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{G})$. A simple way to estimate a GLMM is using the Penalized Quasilikelihood method (PQL) (Schall 1991,

Breslow and Clayton 1993). PQL method can be implemented by iterative fitting a linear mixed model to a modified dependent variable, i.e. it is a method based on iterative fitting of a working linear mixed model. So, the PQL estimates are obtained of the fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{\alpha}$, considering $\boldsymbol{\alpha}$ fixed and penalizing the log-likelihood according to the distribution of $\boldsymbol{\alpha}$, i.e. the estimation of the fixed and random effects coefficients is via Fisher scoring algorithm (Green 1987), as the iterative solution to the system

$$\begin{bmatrix} \boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} & \boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{Z} \\ \boldsymbol{Z}'\tilde{\boldsymbol{W}}\boldsymbol{X} & \boldsymbol{Z}'\tilde{\boldsymbol{W}}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} \\ \boldsymbol{Z}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} \end{bmatrix}, \tag{5.6}$$

where, $\tilde{\boldsymbol{W}}$ and $\tilde{\boldsymbol{z}}$ are defined as in (5.3). Notice that for a GLMM, we have that the matrix $\mathbf{R}^{-1}$ in (1.7) becomes $\boldsymbol{W}$ and the response variable $\boldsymbol{y}$ becomes $\boldsymbol{z}$.

The model components of a P-GLMM are computed analogously to the normal case, a transformation matrix allow us to reparameterize a P-GLM as a P-GLMM (see Section 1.2.2).

Since $\boldsymbol{G}$ depends on the variance component $\sigma^2_{\boldsymbol{\alpha}}$ and $\boldsymbol{W}$ on ϕ (if it is unknown), they are estimated maximizing the REML (Patterson and Thompson 1971), i.e., maximizing (1.10) where $\boldsymbol{R}$ and $\boldsymbol{y}$ are replaced by $\tilde{\boldsymbol{W}}$ and $\tilde{\boldsymbol{z}}$, respectively. Therefore, the PQL solution is achieved by iteration between (5.6) and the solution of the REML maximization.

The PQL solution is the most common estimation procedure for the generalized linear mixed model. However, it is only an approximation to a full likelihood analysis (it is exact in the Gaussian GLMM). It should be taken into account that the PQL tends to underestimate variance components as well as regression coefficients in some situations (e.g. binomial response), since the Laplace approximation methods may be numerically unstable. For those cases, a bias-correction procedure based on higher order Laplace approximation can be used to improve the asymptotic performance of the PQL estimates, see Breslow and Lin (1995) and Breslow and Lin (1996). However, such indirect shemes need not convergence and fail to do so in some practical analysis, in such cases, other methods such as the one proposed in Wood (2011) can be used. Wood (2011) uses a Laplace approximation to obtain an approximate REML which is suitable for efficient direct optimization.

5.2.1 Prediction in P-GLMMs

Analogously to the case of the P-GLM, the methodology associated to a P-GLMM can be extended to predict new out-of-sample values. As it was shown under the Gaussian framework (see Section 2.2.2), to reformulate the extended penalized generalized linear model as a mixed model, an extended transformation matrix is needed. In the case of Gaussian data, we have proved that for a univariate model the fit does not change when the fit and the prediction are obtained simultaneously independently of the extended transformation used. However, this does not happen in the P-GLMM context, since different transformation matrices implies different weight matrices and different working vectors. Therefore, to preserve the fit, the extended model matrices ($\boldsymbol{X}_+$ and $\boldsymbol{Z}_+$) have to be direct extensions of the model matrices used to fit the data ($\boldsymbol{X}$ and $\boldsymbol{Z}$). The transformation matrices that verified the previous condition were already defined in Section 3.3.2.

Once the model components, $\boldsymbol{X}_+$, $\boldsymbol{Z}_+$ and $\boldsymbol{G}_+$, are defined, we propose to fit and predict simultaneously estimating the parameters and coefficients through the PQL method, i.e. the estimation of the extended fixed and random effects ($\boldsymbol{\beta}_{\tilde{+}}$ and $\boldsymbol{\alpha}_+$) and the selection of the random covariance and the dispersion parameters are done iterating between the solution of the extended system of equations and the solution of the extended REML. It means, replacing $\boldsymbol{R}_+^{-1}$ by $\boldsymbol{W}_+$ and $\boldsymbol{y}_+$ by $\boldsymbol{z}_+$ in (2.18) and (2.21), with $\boldsymbol{W}_+$ and $\boldsymbol{z}_+$ defined as in (5.5).

Following the proof given for a univariate model with Gaussian dependent variable, it is straightforward to prove that if the extended model matrices are direct extensions of the model matrices used to fit the data, the fit obtained with the extended components through the PQL method is the same as the solution we obtain only fitting the data (notice that this also implies that the solutions of the REML and the extended REML are the same, i.e. the covariance components estimated fitting and predicting simultaneously are the same as the covariance components estimated only fitting the data). It must be said that other estimation procedures, such as the one introduced in Wood (2011), could also be extended to fit and predict simultaneously, since although they are not based on a working model, the pseudodata, $\boldsymbol{z}_+$, and the weights, $\boldsymbol{W}_+$, could also be defined as in the proposed extended PQL procedure.

5.3 Restrictions for prediction with P-GLMs and P-GLMMs

As in the Gaussian case, simultaneous fit and prediction in models with interaction terms provides changes in the fit. Restrictions to avoid this problem or for other purposes can also be imposed in the contexts of P-GLMs and P-GLMMs through the Lagrange multipliers. In this section, we briefly explain how to impose restrictions under both frameworks.

In a P-GLM, suppose that we have the following restricted extended regression model

$$g(\boldsymbol{y}_+) = \boldsymbol{B}_+\boldsymbol{\theta}^*_+$$

subject to $\boldsymbol{C}\boldsymbol{\theta}^*_+ = \boldsymbol{r}$, where $\boldsymbol{C}$ is a constraint matrix of dimension $l \times c_+$ acting on all coefficients and $\boldsymbol{r}$ is the restrictions vector of dimension $l \times 1$. Notice that we use the superscript $(^*)$ to indicate that we are imposing restrictions.
The Lagrange formulation of the extended penalized log-likelihood (5.4) function is:

$$\mathcal{L}_p(\boldsymbol{\theta}^*_+; \boldsymbol{y}_+) = \boldsymbol{y}'_+\boldsymbol{B}_+\boldsymbol{\theta}^*_{n_+} - \mathbf{1}'_+ b(\boldsymbol{B}_+\boldsymbol{\theta}^*_+) + \mathbf{1}'_{n_+} c(\boldsymbol{y}_+) - \tfrac{1}{2}\lambda\boldsymbol{\theta}'^*_+\boldsymbol{P}_+\boldsymbol{\theta}^*_+ - \boldsymbol{\omega}'(\boldsymbol{C}\boldsymbol{\theta}^*_+ - \boldsymbol{r}), \quad (5.7)$$

$\boldsymbol{\theta}^*_+$ denotes the restricted least squares (RLS) estimator and $\boldsymbol{\omega}$ is a $l \times 1$ vector of Lagrangean multipliers. Differentiating (5.7) we find

$$\begin{aligned} \frac{\partial \mathcal{L}_p}{\partial \boldsymbol{\theta}^*_+} &= \boldsymbol{B}'_+\boldsymbol{y}_+ - \boldsymbol{B}'_+\boldsymbol{\mu} - \lambda\boldsymbol{P}_+\boldsymbol{\theta}_+ - \boldsymbol{C}'\boldsymbol{\omega}, \\ \frac{\partial \mathcal{L}_p}{\partial \boldsymbol{\omega}} &= -\boldsymbol{C}\boldsymbol{\theta}^*_+ + \boldsymbol{r}. \end{aligned}$$

The previous system is nonlinear, our proposal to solve it is to solve the following iterative re-weighted system

$$\begin{bmatrix} \boldsymbol{B}'_+\tilde{\boldsymbol{W}}_+\boldsymbol{B}_+ + \lambda\boldsymbol{P}_+ & \boldsymbol{C}' \\ \boldsymbol{C} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}}^*_+ \\ \hat{\boldsymbol{\omega}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}'_+\tilde{\boldsymbol{W}}_+\tilde{\boldsymbol{z}}_+ \\ \boldsymbol{r} \end{bmatrix}.$$

The restricted estimating PIRLS algorithm is summarized as follows:

---

1. set an initial value of $\boldsymbol{\theta}^*_+$ $(= \hat{\boldsymbol{\theta}}^*_{+_{old}})$
2. use $\hat{\boldsymbol{\theta}}^*_{+_{old}}$ to estimate $\boldsymbol{W}_+$, $\boldsymbol{\mu}_{old}$ and $\boldsymbol{\omega}_{old}$
3. let $\hat{\boldsymbol{\eta}}_{+_{old}} = \boldsymbol{B}_+\hat{\boldsymbol{\theta}}^*_{+_{old}}$, get the $\boldsymbol{z}_{+_{new}}$
4. obtain the new estimate $\hat{\boldsymbol{\theta}}^*_{+new}$

5. repeat 2 to 4 until the convergence criterion is satisfied $\left(\text{e.g. } \frac{\sum_{i=1}^{n}(\eta_{+new_i}-\eta_{+old_i})^2}{\sum_{i=1}^{n}\eta^2_{+new_i}} < \delta\right)$ for a small value of δ ($\delta \simeq 10^{-6}$)

It means, everything is analogous to the Gaussian case substituting $\boldsymbol{R}_+^{-1}$ by $\tilde{\boldsymbol{W}}_+$ and $\boldsymbol{y}_+$ by $\tilde{\boldsymbol{z}}_+$ in (3.25).

Under the P-GLMM framework, the imposition of the restrictions is also analogous to the Gaussian case. Suppose that we consider the following restricted extended generalized linear mixed model:

$$g(\boldsymbol{y}_+) = \boldsymbol{X}_+\boldsymbol{\beta}^*_{\tilde{+}} + \boldsymbol{Z}_+\boldsymbol{\alpha}^*_+$$

subject to $\boldsymbol{C}_{\text{MM}}\begin{bmatrix}\boldsymbol{\beta}^*_{\tilde{+}}\\ \boldsymbol{\alpha}^*_+\end{bmatrix} = \boldsymbol{r}_{\text{MM}}$, where $\boldsymbol{\alpha}^*_+ \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}_+)$, $\boldsymbol{C}_{\text{MM}}$ is a constraint matrix of dimension $l \times c_+$ acting on all coefficients and $\boldsymbol{r}_{\text{MM}}$ is the restrictions vector of dimension $l \times 1$.
Thus, for given values of the variance components, the restricted coefficients are obtained by maximizing the restricted penalized log-likelihood:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}^*_{\tilde{+}}, \boldsymbol{\alpha}^*_+, \boldsymbol{w}) &= \boldsymbol{y}'_+(\boldsymbol{X}_+\boldsymbol{\beta}^*_{\tilde{+}} + \boldsymbol{Z}_+\boldsymbol{\alpha}^*_+) - \boldsymbol{1}'_{n_+}b(\boldsymbol{X}_+\boldsymbol{\beta}^*_{\tilde{+}} + \boldsymbol{Z}_+\boldsymbol{\alpha}^*_+) + \boldsymbol{1}'_{n_+}c(\boldsymbol{y}_+)\\ &\quad -\frac{1}{2}\boldsymbol{\alpha}'^*_+\boldsymbol{G}_+^{-1}\boldsymbol{\alpha}^*_+ - \boldsymbol{\omega}'\left(\boldsymbol{C}_{MM}\begin{bmatrix}\boldsymbol{\beta}^*_{\tilde{+}}\\ \boldsymbol{\alpha}^*_+\end{bmatrix} - \boldsymbol{r}_{MM}\right).\end{aligned}$$

Defining $\boldsymbol{C}_{MM} = \left[\boldsymbol{C}_{MM_f} \mid \boldsymbol{C}_{MM_r}\right]$, the first order conditions yield

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}^*_{\tilde{+}}} &= \boldsymbol{X}'_+\boldsymbol{y}_+ - \boldsymbol{X}'_+\boldsymbol{\mu}^*_+ - \boldsymbol{C}'_{MM_f}\boldsymbol{\omega}\\ \frac{\partial\mathcal{L}}{\partial\boldsymbol{\alpha}^*_+} &= \boldsymbol{Z}'_+\boldsymbol{y}_+ - \boldsymbol{Z}'_+\boldsymbol{\mu}^*_+ - \boldsymbol{G}_+^{-1}\boldsymbol{\alpha}_+ - \boldsymbol{C}'_{MM_r}\boldsymbol{\omega}\\ \frac{\partial\mathcal{L}}{\partial\boldsymbol{w}} &= \boldsymbol{C}_{MM_f}\boldsymbol{\beta}^*_{\tilde{+}} + \boldsymbol{C}_{MM_r}\boldsymbol{\alpha}^*_+ - \boldsymbol{r}_{MM}\end{aligned}$$

Our proposal to solve the previous system of equations is to solve the following iterative re-weighted system

$$\begin{bmatrix}\boldsymbol{X}'_+\tilde{\boldsymbol{W}}_+\boldsymbol{X}_+ & \boldsymbol{X}'_+\tilde{\boldsymbol{W}}_+\boldsymbol{Z}_+ & \boldsymbol{C}'_{MM_f}\\ \boldsymbol{Z}'_+\tilde{\boldsymbol{W}}_+\boldsymbol{X}_+ & \boldsymbol{Z}'_+\tilde{\boldsymbol{W}}_+\boldsymbol{Z}_+ + \boldsymbol{G}_+^{-1} & \boldsymbol{C}'_{MM_r}\\ \boldsymbol{C}_{MM_f} & \boldsymbol{C}_{MM_r} & \boldsymbol{O}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}^*_{\tilde{+}}\\ \boldsymbol{\alpha}^*_+\\ \boldsymbol{\omega}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}'_+\tilde{\boldsymbol{W}}_+\tilde{\boldsymbol{z}}_+\\ \boldsymbol{Z}'_+\tilde{\boldsymbol{W}}_+\tilde{\boldsymbol{z}}_+\\ \boldsymbol{r}_{MM}\end{bmatrix} \tag{5.8}$$

with $\tilde{\boldsymbol{W}}_+$ and $\tilde{\boldsymbol{z}}_+$ defined as in (5.5). The restricted estimating PIRLS used is analogous

to the one explained above for the P-GLM framework. If the restriction is the fit is maintained, the covariance parameters are the ones estimated in the fit. Otherwise, these parameters are estimated through the extended REML, i.e., the solution of the previous problem is achieved by iteration between (5.8) and the solution of the associated extended REML maximization problem.

5.4 Application

To illustrate the methodology introduced in this chapter, we analyze American data on the incidence of respiratory disease analyzed in Currie et al. (2006). Among other information, the dataset contains the number of deaths according to the age at death (ranging from 40 to 100), the calendar year of death (from 1959 to 1998), and the month of death (from 1 to 12). We have selected the year 1990, and considered as regressors the months of death and the ages of death. In order to predict and compare the results with the known values we have divided the dataset into two groups: an observed dataset that contains the number of deaths for ages ranging from 40 to 90 (612 values) and a training dataset that contains the number of deaths for ages ranging from 91 to 100 (120 values). The fit and the prediction was obtained through four different models:

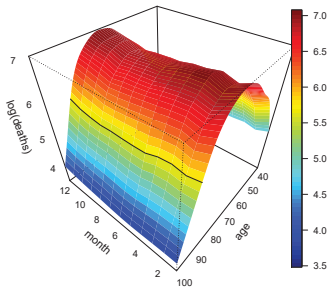
a) Model 1, 2D P-spline model: $g(\boldsymbol{y}_+) = f(\boldsymbol{month}, \boldsymbol{age}_+)$ where $\boldsymbol{y}_+$ are the death counts.

b) Model 2: The model defined in a) subject to the restrictions the fit is maintained.

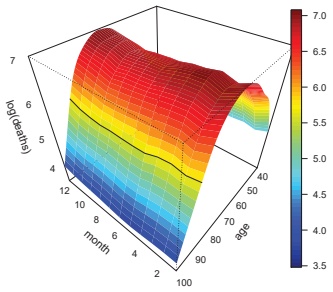
c) Model 3, S-ANOVA model: $g(\boldsymbol{y}_+) = f_1(\boldsymbol{month}) + f_2(\boldsymbol{age}_+) + f_{1,2}(\boldsymbol{month}, \boldsymbol{age}_+)$ where $\boldsymbol{y}_+$ are the death counts.

d) Model 4: The model defined in c) subject to the restriction the fit is maintained.

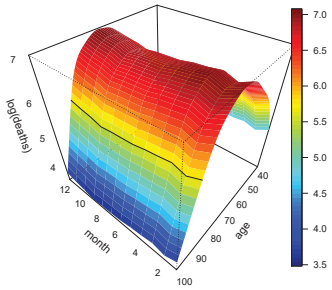
Figure 5.1 shows the fit and the prediction obtained with model 1, model 2, model 3 and model 4.
As can be appreciated the four models provide almost the same solution. In order to illustrate this fact, we have selected months 3 and 9. Figure 5.2 shows the fit, the prediction and the associated 95% confidence intervals obtained through the four models for months 3 (left panel) and 9 (right panel), the solutions are almost equal, however, the prediction performances of the S-ANOVA models are a bit better that of the 2D P-spline models for month 9. The differences between the restricted and unrestricted models are not significant.



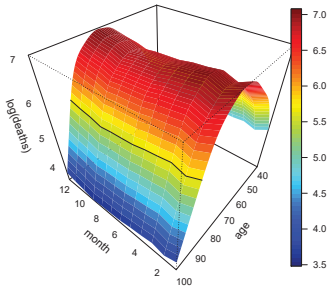
(a) Model 1.



(b) Model 2.



(c) Model 3.



(d) Model 4.

Figure 5.1: Fit and prediction of a data set on death counts of American males, from ages 40 to 100 over the months $1 - 12$, through model 1, model 2, model 3 and model 4, from top left to bottom right respectively.

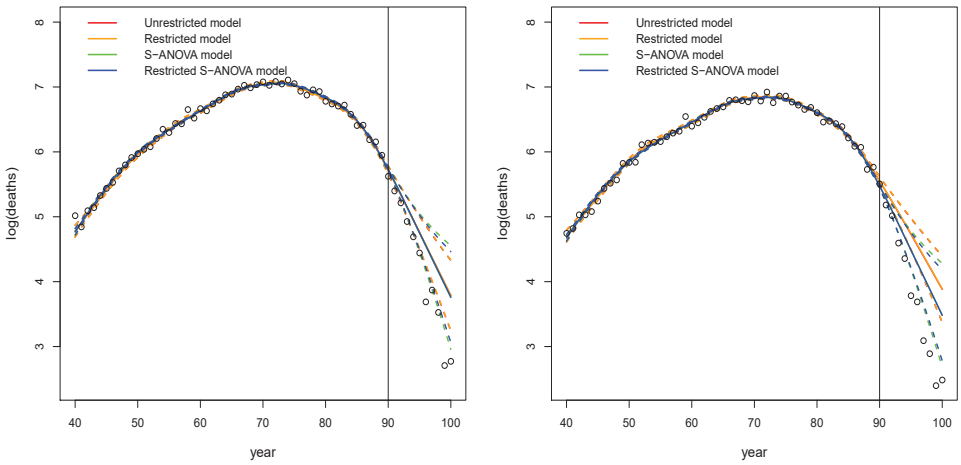


Figure 5.2: Fit and prediction of selected months 3 (left panel) and 9 (right panel) obtained through model 1 (red line), model 2 (orange line), model 3 (green line) and model 4 (blue line). The dashed lines are the 95% confidence intervals. The vertical line indicates the age from which we predict (90).

5.5 Summary of the chapter

In this chapter we have extended the prediction methodology for a generalized linear model in the context of P-splines and mixed models.

To generalize the methodology, we have used the same idea that estimation procedures use to solve the nonlinear equations that we face in the GLMs context: use the associated working normal theory model or the pseudodata $\boldsymbol{z}$ and weights $\boldsymbol{W}$ and follow the procedure used under under the assumption of Gaussianity.

Based on the previous idea, we have extended the algorithms used in the estimation procedures to fit and predict simultaneously with P-GLMs and P-GLMMs and to impose restrictions over the coefficients. In the context of mixed models, we have warned about the importance of using a transformation that preserves the original model matrices to maintain the fit, even in the univariate case, since different transformations deal with different working vectors and therefore with different solutions.

The methodology has also been illustrated with one real data example to predict deaths due to respiratory disease.

Chapter 6

Conclusions and further work

Summary of contributions of the thesis

Statistical methods for the smooth of data have found many different applications along the years. One challenge is to extend these techniques to predict out-of-sample values. In this dissertation we have based our research on methods available in the literature to deepen its study. The objective has been to propose a general framework for out-of-sample prediction. In this dissertation, we have considered semi-parametric smoothing techniques for the prediction of out-of-sample observations. The proposed methodology could be used for any regression basis and quadratic penalty. However, we have focused our work on penalized splines models (B-splines basis with discrete penalties) as an unified framework to predict out-of-sample values and to incorporate restrictions for coherent predictions. All the methodology related to P-splines and its reparameterization as a mixed model has been reviewed in Chapter 1.

In Chapter 2, we have given a general approach for out-of-sample prediction with smooth models that do not include interaction terms. To predict in the framework of penalized regression with any regression basis and any quadratic penalty, we have generalized methods available in different statistical areas: penalized regression (Currie et al. 2004), mixed models (Gilmour et al. 2004) and global optimization (Sacks et al. 1989). For the particular case of penalties based on differences between adjacent coefficients, we have proved the equivalence of all the methods. Also, we have shown that the prediction does not influence the fit to the data (note that this implies that the variance components obtained through the REML and the extended REML are the same).
It is worth noting that, since all methods provide the same solution, out-of-sample prediction can be done through one or two steps procedures. The advantage of a one-stage procedure is obviously that the fit and the prediction are obtained simultaneously. However, two-stage approaches allow us to compute prediction intervals based on the conditional

distribution of $\boldsymbol{y}_p|\boldsymbol{y}$ (i.e. prediction intervals that are not obtained through the common definition of prediction intervals, they are computed from a conditioned distribution). Moreover, we have provided valuable insights into the relationships between the penalty order and the shape of the prediction. The shape of the prediction is determined by the penalty order, for instance, for penalty order one the prediction is constant and for penalty order two the prediction is linear. Although linear prediction is the most common shape, other shapes can be obtained simply by changing the penalty order.

In areas such as Demography, when mortality tables are forecasted, we might be interested in knowing which known data we are using to obtain the prediction, or in using a subset of the dataset to predict. For such purposes, we have introduced the concept of "memory of a P-spline" as a tool to know how much known information we use to predict. Our definition is just one possible way to determine the "memory of a P-spline", and hence other summary statistics could be used. The properties of the concept have been obtained and proved through a simulation study, being the most important one that the amount of smoothness determines the "memory of a P-spline". Hence, this term allow us to know which observations influence in the prediction, or, perhaps more interestingly, to compute the amount of smoothness for a given value of the memory.

Through several examples, we have shown the potential of the methods proposed: i) the analysis of aboveground biomass shows that our approach allows to predict to both sides to the data (to the left and to the right); ii) other dataset on monthly sulphur dioxide levels illustrates that seasonal components can be taken into account in the prediction, in order to do that we have used the smooth modulation model based on P-splines (Eilers et al. 2008); and iii) we show how out-of-sample prediction can be performed in the case of correlated errors with the analysis of a dataset on yearly sea level. In this case, we have provided the implementation of the algorithm that allow us to compute the variance and correlation components.

In the third chapter, we have provided a method for out-of-sample prediction in the case of multidimensional smoothing. The results are not a straightforward extension of the results in Chapter 2, since out-of-sample prediction influence the fit to the data due to the structure introduced by the Kronecker products. In models without and with interaction, the extended penalty matrix is not a direct extension of the penalty matrix used to obtain the fit. However, as we have shown in Chapter 2, in models without interaction terms, the blocks of the extended penalty matrix simplify and the fit is not altered. We have shown that this does not occur in models with interaction terms.

When out-of-sample prediction is carried out just in one covariate, we have proved which are the coefficients from the fit that determine the prediction, and that the penalty can be modified in order to preserve the fit. However, if we modify the penalty we do not impose the penalty correctly and the argument cannot be extended when prediction has to be carried out in more than one dimension. As an unified solution to preserve the fit, we have proposed to impose restrictions over the coefficients through the use of Lagrange multipliers. A proposal, which may be useful for other purposes, such as incorporating known information about the unobserved values.

The prediction methodology using Lagrange multipliers was introduced in the P-splines and mixed models framework. In the P-splines framework, the imposition of restrictions does not imply major difficulties. However, in the mixed models context, we have shown that to impose restrictions without taking into account the model reparameterization as a P-spline, attention needs to be paid since the matrices of fixed and random effects for out-of-sample prediction need to be direct extensions of the matrices used in the fit. To that end, we have defined an extended transformation matrix that preserves the model matrices used in the fit.

To illustrate the methodology and test the usefulness of the restrictions, we have solved a crossover problem in a real mortality dataset by imposing restrictions over the coefficients to maintain the fit and in order to avoid higher mortality rates for younger ages than for older ages. We also compared the results with the model developed in Delwarde et al. (2007), since they have improved the Lee-Carter, and this is one of the most common methods used for estimating and forecasting mortality data.

An efficient approach to include smooth additive terms and smooth interactions are the so-called Smooth-ANOVA models. They allow us to include interaction terms that can be decomposed as a sum of several smooth functions. In Chapter 4, we extended the prediction methodology developed in Chapter 3 to this modelling framework. These models provide a very general framework for data analysis (e.g. including higher order interactions), and therefore, it would be interesting to use them for out-of-sample prediction. However, they suffer from identifiability problems, to overcome this problem we have followed Lee and Durbán (2011) and reparameterized them as mixed models.

We have seen that in order to write a S-ANOVA model as a mixed model, we need to know the B-spline basis and the penalty matrix, but once the model is written as a mixed

model we can not formulate it as a P-spline, because the transformation matrix used to avoid identifiability problems is not orthogonal. This is why for the invariance of the fit in a S-ANOVA model, an extended transformation matrix that preserves the model matrices is indispensable, we have defined such matrix in Section 4.2.2.

The present work has illustrated the prediction with Smooth-ANOVA models by analyzing the aboveground biomass dataset, the Smooth-ANOVA model allows to represent the smooth function as the sum of a smooth function for the height, a smooth function for the diameter and a smooth term for the height-diameter interaction. Further, we have done a simulation study to evaluate the accuracy of the 2D interaction P-spline models and Smooth-ANOVA models, both models with and without the restriction the fit has to be maintained. From the results of the simulation study, we have concluded that the imposition of invariance in the fit can improve the fit and the prediction, and that in most situations the constrained S-ANOVA model behaves better in the fit and out-of-sample predictions, however results depend on the simulation scenario and on the dimensions the prediction is carried out (one or both dimensions).

The fifth chapter was devoted to extend the methodology developed in the previous chapters for generalized linear models (GLMs) in the context of P-splines (P-GLMs) and mixed models (P-GLMMs). GLMs are probably the most important models in statistics. A vast majority of models are just special cases of a GLM, or generalizations of it. Therefore, an out-of-sample prediction strategy in the context of P-GLMs is potentially very useful in a wide range of applications. In order to develop such strategy, we have extended one of the coefficients and parameters estimation procedures used in the context of P-GLMMs to fit and predict simultaneously, the Penalized Quasilikelihood method (PQL). The iterative algorithm used in the PQL estimation procedure is based on a working normal theory model. Analogously to the Gaussian case, we have proposed to give infinite variance to the unknown observations. We have pointed out the importance of using an extended transformation matrix that preserves the model matrices, even in the univariate case, since different transformations deal with different working vectors and therefore with different solutions. The present work has also showed how restrictions can be imposed in P-GLMs and P-GLMMs models.

Through 2D interaction P-splines and S-ANOVA models we have illustrated our proposal to predict out-of-sample values in the context of GLMs (imposing and without imposing restrictions). From a dataset on deaths due to respiratory disease for ages ranging from 40 to 90, we have predicted the deaths for ages between 91 to 100. By dividing the

dataset into two (a training and a testing dataset), we have been able to compared the predicted values with the real ones, and to conclude that the models are quite accurate, since the real data follow the trends predicted by the models.

Further research

The research carried out in this thesis has highlighted some questions areas that could be extended or improved, and therefore, require further research:

i) Out-of-sample prediction in the case of complex and/or non-discrete penalties.

Based on the work done in Chapters 2 and 3, we see that penalties play an important role on predictions, and notice that they are not invariant (they change when the fit and out-of-sample prediction is carried out simultaneously) in the case of multidimensional smoothing. Rather than constraining the prediction one could attempt to develop new penalties to enforce desired properties, specially in the context of continuous penalties based on differential equations (Ramsay et al. 2016).

ii) Memory of a P-spline as a smoothing criteria.

It is well known the fact that out-of-sample prediction, when using P-splines, is driven by the trend present in the last observations. For example, in the case of mortality data, the mortality rates in the last years will have a strong impact on the prediction of future mortality. This can be a problem if, for whatever reason the trend over the recent past is quite different from the overall trend in the data. A researcher or practitioner may decide to carry out the out-of-sample prediction, not conditioned by the optimal amount of smoothing, but deciding the number of years that should influence the prediction. Further research is needed to provide optimal selection criteria based on the memory of a P-splines, as well as extending this concept to the case of multidimensional smoothing.

iii) Out-of-sample and multiscale prediction.

In areas such as Demography, Public Health or Epidemiology, aggregated data frequently appear due to patients' confidentiality or compact presentation. For instance, death counts are commonly recorded or grouped by age classes, year intervals and/or geographical units, but the interest might be to obtain predictions at a resolution different from the original one (for example, death count estimates by single age, calendar year and/or smaller nested units than the original ones). Ayma et al. (2016a) and Ayma et al. (2016b) offer a flexible methodology to deal with in-sample prediction at different spatial and temporal scales. Thus, it would

be interesting to extend these works to include also out-of-sample prediction. This extension will allow us to simultaneously estimate the underlying distribution behind aggregated data and predict in a more detailed temporal scale in out of the available data sample, all of this under a GLM (or GLMM) context.

iv) Non-linear constraints in out-of-sample prediction.

In Chapter 3, we have shown that the imposition of restrictions over the coefficients is potentially very useful in order to maintain the fit or to avoid crossover problems. But more interesting restrictions could be imposed, for instance, in mortality forecast we may want to impose that closely related subpopulations have non-divergent mortality trends over time. In particular, male mortality is larger than female mortality at the same age, however, if male and female mortality are projected separately it could happen that in the future male mortality rates become lower than female ones. Restrictions can be imposed to overcome this problem or to ensure that the predicted mortality values is above or below a threshold. Further work is still needed to impose nonlinear restrictions, this would imply combining the methods proposed with optimization techniques.

v) Out-of-sample prediction under different estimation procedures.

As we have pointed in Chapter 5, our approach to out-of-sample prediction in the context of P-GLMs and P-GLMMs can be extended for any estimation procedure based on a working model or a set of pseudodata and weights. From a statistical point of view, it would be interesting to provide the software for the implementation of algorithms to fit and predict simultaneously in the context of GLMs following our approach given in Chapter 5, but using estimation procedures that do not suffer from bias problems (as it happens with PQL for binary data), such as the one developed by Wood (2011).

References

Ayma, D., Durban, M., Lee, D.-J., and Eilers, P. (2016a). Penalized composite link models for aggregated count data: a mixed model approach. *Spatial Statistics*, 17:179–198.

Ayma, D., Durban, M., Lee, D.-J., and Kassteele, J. (2016b). Modelling latent trends from spatio-temporally grouped data using composite link mixed models. *UC3M Working Papers Statistics and Econometrics*, 16-07.

Ba, A., Sinn, M., Goude, Y., and Pompey, P. (2012). Adaptive learning of smoothing functions: Application to electricity load forecasting. In Pereira, F., Burges, C., Bottou, L., and Weinberger., K., editors, *Advances in Neural Information Processing Systems 25*.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic system. *Statistical Science*, 10(1):3–66.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Breslow, N. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, 82:81–91.

Breslow, N. and Lin, X. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association, Biometrika*, 91(435):1007–1016.

Camarda, C. (2012). Mortalitysmooth: An R package for smoothing Poisson counts with P-splines. *Journal of Statistical Software*, 50(I).

Cressie, N. (1993). *Statistics for Spatial data*. Wiley: New York.

Currie, I. (2013). Smoothing constrained generalized linear models with an application to the Lee-Carter model. *Statistical Modelling*, 13(1):69–93.

Currie, I. and Durbán, M. (2002). Flexible smoothing with P-splines: A unified approach. *Statistical Modelling*, 2:333–349.

Currie, I., Durbán, M., and Eilers, P. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, 4(4):279–298.

Currie, I., Durbán, M., and Eilers, P. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statististical Society, Series B*, 68:1–22.

Delwarde, A., Denuit, M., and Eilers, P. (2007). Smoothing the Lee-Carter and Poisson log-bilinear models for mortality forecasting: a penalized log-likelihood approach. *Statistics and Modelling*, 7:29–48.

Eilers, P., Gampe, J., Marx, B., and Rau, R. (2008). Modulation models for seasonal time series and incidence tables. *Statistics in Medicine*, 27:3430–3441.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Eilers, P. and Marx, B. (2010). Splines, knots, and penalties. *Computational Statistics*, 2(6):637–653.

Eilers, P., Marx, B., and Durbán, M. (2015). Twenty years of P-splines. *Statistics and Operations Research Transactions*, 39(2):149–186.

Etxeberria, J., Ugarte, M., Goicoa, T., and Militino, A. (2015). On predicting cancer mortality using ANOVA-type P-splines models. *Revstat - Statistical Journal*, 13(1):21–40.

Gilmour, A., Cullis, B., Welham, S., Gogel, B., and Thompson, R. (2004). An efficient computing strategy for prediction in mixed linear models. *Computational Statistics and Data Analysis*, 44:571–586.

Girosi, F., Jones, M., and Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269.

Green, P. (1987). Penalized likelihod for general semi-parametric regression models. *International Statistical Review*, 55(3):245–259.

Greene, W. and Seaks, G. (1991). The restricted least squares estimator: A pedagogical note. *The Review of Economics and Statistics*, 73(3):563–567.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–338.

Harville, D. (2000). *Matrix Algebra from a Statistician's Perspective*. Springer.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall.

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423–447.

Human Mortality Database (2018). University of california, berkeley (usa), and max planck institute for demographic research (germany). https://mortality.org.

Hyndman, R.J. and. Koehler, A. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 4(22):679 – 688.

Hyndman, R., Koehler, A., Ord, J., and Snyder, R. (2008). *Forecasting with Exponential Smoothing*. Springer Series in Statistics.

Jones, D., Schonlau, M., and William, J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimiztion*, 13:455–492.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of computational and graphical statistics*, 13(1):183–212.

Lee, D.-J. (2010). *Smoothing mixed models for spatial and spatio-temporal data*. PhD thesis, Department of Statistics, Universidad Carlos III de Madrid, Spain.

Lee, D.-J. and Durbán, M. (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, 11(1):49–69.

Lee, D.-J. and Durbán, M. (2012). Seasonal modulation smoothing mixed models for time series forecasting. In *Proceedings of 27th International Workshop on Statistical Modelling*, Prague, Czech Republic.

Lee, R. and Carter, L. (1992). Modelling and forecasting the time series of US mortality. *Journal of the American Statistical Association*, 87:659–71.

Marx, B. and Eilers, P. (1999). Generalized linear regression on sampled signals and curves: A p-spline approach. *Technometrics*, 41:1–13.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, New York.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, 9(2):363–379.

Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proceedings of IEEE*, 78:1481–1497.

Pollice, A. and Bilancia, M. (2002). Kriging with mixed effects models. *Statistica*, 3:405–429.

Ramsay, J., Hooker, G., Campbell, D., and Cao, J. (2016). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:741–796.

Rivas-Martínez, D., Díaz, T., Fernández-González, F., Izco, J., Loidi, J., Lousã, M., and Penas, A. (2002). Vascular plant communities of spain and portugal. addenda to the syntaxonomical checklist of 2001. *Itinera Geobotanica*, pages 15, 1–2, 5–22.

Rodríguez-Álvarez, M., Lee, D.-J., Kneib, T., Durbán, M., and Eilers, P. (2018). On the estimation of variance parameters in non-standard generalised linear mixed models: application to penalised smoothing. *Statistics and Computing*, 29:1–18.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Series in Statistical and Probabilistic Mathematics. Cambridge University Press, UK.

Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–435.

Sánchez-González, M., Durbán, M., Lee, D., Cañellas, I., and Sixto, H. (2016). Smooth additive mixed models for predicting aboveground biomass. *Journal of Agricultural, Biological and Environmental Statistics*, 22:23–41.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–721.

Searle, S., Casella, G., and McCulloch, C. (1992). *Variance components*. Wiley Series in Probability and Mathematical Statistics.

Ugarte, M., Goicoa, T., Etxeberria, J., and Militino, A. (2012). Projections of cancer mortality risks using spatio-temporal P-spline models. *Statistical Methods in Medical Research*, 21(5):545–560.

Wand, M. (2003). Smoothing and mixed models. *Computational statistics*, 18:223–249.

Wood, S. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B*, 73:3–36.

Yi, G., Shi, J., and Choi, T. (2011). Penalized Gaussian Process Regression and Classification for High-Dimensional Nonlinear Data. *Biometrics*, 6:1285–1294.

Appendix A

Appendix to Chapter 2

A.1 R code to extend the basis matrix

In this section we show the code to build a new extended B-spline basis, $\boldsymbol{B}_+$, built from a new set of knots that contains the knots used to build the original basis $\boldsymbol{B}$.

```
# Extended matrix B_+ given the degree, bdeg, the knots used to fit the data, knots,
# and the extended covariate, x.extended
 library(splines)
 end = max(x.extended)
 dx = diff(knots)
 knots.aux = seq(from = max(knots) + dx, to = end + 100*dx, by = dx)
 knots.aux2 = c(knots.aux[1:which(knots.aux>=end)[1]-1],knots.aux[which(knots.aux>=end)[1:bdeg]])
 knots.ext = c(knots, knots.aux2)
 B.extended = spline.des(knots.ext, x.extended, bdeg + 1, sparse=FALSE, outer.ok=TRUE)$design
```

A.2 Derivatives of the approximate restricted maximum likelihood with respect to the variance and correlation components

In this section we present the derivatives of the approximate restricted maximum likelihood with respect to the variance and correlation components.

Given the approximate log-likelihood

$$l(\sigma^2_{\boldsymbol{\epsilon}}, \sigma^2_{\boldsymbol{\alpha}}, \rho) = \underbrace{-\frac{1}{2}\log|\boldsymbol{V}|}_{\text{Part I}} \underbrace{-\frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|}_{\text{Part II}} \underbrace{-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)}_{\text{Part III}},$$

with $\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{ZGZ}'$, where $\boldsymbol{R} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}$ and $\boldsymbol{G} = \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{I}$. The derivatives with respect to the variance and correlation components are the following.

Estimation of $\sigma^2_{\boldsymbol{\epsilon}}$.
Derivative of Part I with respect to $\sigma^2_{\boldsymbol{\epsilon}}$:

$$\begin{aligned} \frac{\partial\left(\frac{1}{2}\log|\boldsymbol{V}|\right)}{\partial\sigma^2_{\boldsymbol{\epsilon}}} &= \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\sigma^2_{\boldsymbol{\epsilon}}}\right) \\ &= \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\tilde{\boldsymbol{R}}\right). \end{aligned}$$

Derivative of Part II with respect to $\sigma^2_{\boldsymbol{\epsilon}}$:

$$\begin{aligned} \frac{\partial\left(\frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|\right)}{\partial\sigma^2_{\boldsymbol{\epsilon}}} &= \frac{1}{2}\text{trace}\left(\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\frac{\partial\boldsymbol{V}^{-1}}{\partial\sigma^2_{\boldsymbol{\epsilon}}}\boldsymbol{X}\right) \\ &= -\frac{1}{2}\text{trace}\left(\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\sigma^2_{\boldsymbol{\epsilon}}}\boldsymbol{V}^{-1}\boldsymbol{X}\right) \\ &= -\frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\sigma^2_{\boldsymbol{\epsilon}}}\right) \\ &= -\frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\tilde{\boldsymbol{R}}\right). \end{aligned}$$

Derivative of Part III with respect to $\sigma^2_{\boldsymbol{\epsilon}}$:

$$\begin{aligned} \frac{\partial\left(\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\right)}{\partial\sigma^2_{\boldsymbol{\epsilon}}} &= \frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\frac{\partial\boldsymbol{V}^{-1}}{\partial\sigma^2_{\boldsymbol{\epsilon}}}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) \\ &= -\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\sigma^2_{\boldsymbol{\epsilon}}}\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) \\ &= -\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\tilde{\boldsymbol{R}}\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right). \end{aligned}$$

Joining the derivatives of Part I and Part II:

$$\begin{aligned} \frac{\partial\left(-\frac{1}{2}\log|\boldsymbol{V}|-\frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|\right)}{\partial\sigma^2_{\boldsymbol{\epsilon}}} &= -\frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\tilde{\boldsymbol{R}}\right)+\frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\tilde{\boldsymbol{R}}\right) \\ &= -\frac{1}{2}\text{trace}\left(\left(\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\right)\tilde{\boldsymbol{R}}\right) \\ &= -\frac{1}{2}\text{trace}\left(\boldsymbol{P}\tilde{\boldsymbol{R}}\right), \end{aligned}$$

with

$$\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}. \tag{A.1}$$

Therefore,

$$\frac{\partial l}{\partial \sigma^2_{\boldsymbol{\epsilon}}} = -\frac{1}{2}\text{trace}\left(\boldsymbol{P}\tilde{\boldsymbol{R}}\right) + \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\tilde{\boldsymbol{R}}\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right).$$

By equation 5.2 of Harville (1977) we have that $\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{R}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}\right)$, therefore:

$$\begin{aligned}
\frac{\partial l}{\partial \sigma^2_{\boldsymbol{\epsilon}}} &= -\frac{1}{2}\text{trace}\left(\boldsymbol{P}\tilde{\boldsymbol{R}}\right) + \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}\right)'\boldsymbol{R}^{-1}\tilde{\boldsymbol{R}}\boldsymbol{R}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}\right) \\
&= -\frac{1}{2\sigma^2_{\boldsymbol{\epsilon}}}\text{trace}\left(\boldsymbol{P}\boldsymbol{R}\right) + \frac{1}{2\sigma^2_{\boldsymbol{\epsilon}}}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}\right)'\boldsymbol{R}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}\right)
\end{aligned}$$

Since $\boldsymbol{R} = \sigma^2_{\boldsymbol{\epsilon}}\tilde{\boldsymbol{R}}$, we have that $\boldsymbol{R}^{-1} = \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}\tilde{\boldsymbol{R}}^{-1}$, hence:

$$\frac{\partial l}{\partial \sigma^2_{\boldsymbol{\epsilon}}} = -\frac{1}{2\sigma^2_{\boldsymbol{\epsilon}}}\text{trace}\left(\boldsymbol{P}\boldsymbol{R}\right) + \frac{1}{2\sigma^4_{\boldsymbol{\epsilon}}}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}\right)'\tilde{\boldsymbol{R}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}\right),$$

and:

$$\begin{aligned}
-2\frac{\partial l}{\partial \sigma^2_{\boldsymbol{\epsilon}}} = 0 \quad &\Leftrightarrow \quad -\frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}\text{trace}\left(\boldsymbol{P}\boldsymbol{R}\right) + \frac{1}{\sigma^4_{\boldsymbol{\epsilon}}}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}\right)'\tilde{\boldsymbol{R}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}\right) \\
&\Leftrightarrow \quad \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}\text{trace}\left(\boldsymbol{P}\boldsymbol{R}\right) = \frac{1}{\sigma^4_{\boldsymbol{\epsilon}}}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}\right)'\tilde{\boldsymbol{R}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}\right) \\
&\Leftrightarrow \quad \hat{\sigma}^2_{\boldsymbol{\epsilon}} = \frac{\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}\right)'\tilde{\boldsymbol{R}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\alpha}}\right)}{\text{trace}\left(\boldsymbol{P}\boldsymbol{R}\right)}.
\end{aligned}$$

Estimation of $\sigma^2_{\boldsymbol{\alpha}}$.

Derivative of Part I with respect to $\sigma^2_{\boldsymbol{\alpha}}$:

$$\begin{aligned}
\frac{\partial\left(\frac{1}{2}\log|\boldsymbol{V}|\right)}{\partial \sigma^2_{\boldsymbol{\alpha}}} &= \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \sigma^2_{\boldsymbol{\alpha}}}\right) \\
&= \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial \boldsymbol{G}}{\partial \sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'\right).
\end{aligned}$$

Derivative of Part II with respect to $\sigma^2_{\boldsymbol{\alpha}}$:

$$\begin{aligned}
\frac{\partial\left(\frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|\right)}{\partial\sigma^2_{\boldsymbol{\alpha}}} &= \frac{1}{2}\text{trace}\left(\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\frac{\partial\boldsymbol{V}^{-1}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{X}\right) \\
&= -\frac{1}{2}\text{trace}\left(\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{V}^{-1}\boldsymbol{X}\right) \\
&= -\frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\right) \\
&= -\frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'\right).
\end{aligned}$$

Derivative of Part III with respect to $\sigma^2_{\boldsymbol{\alpha}}$:

$$\begin{aligned}
\frac{\partial\left(\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\right)}{\partial\sigma^2_{\boldsymbol{\alpha}}} &= \frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\frac{\partial\boldsymbol{V}^{-1}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) \\
&= -\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) \\
&= -\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) \\
&= -\frac{1}{2}\hat{\boldsymbol{\alpha}}'\boldsymbol{G}^{-1}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{G}^{-1}\hat{\boldsymbol{\alpha}},
\end{aligned}$$

since $\hat{\boldsymbol{\alpha}} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$.

Joining the derivatives with respect to $\sigma^2_{\boldsymbol{\alpha}}$ of Part I and Part II:

$$\begin{aligned}
\frac{\partial\left(-\frac{1}{2}\log|\boldsymbol{V}|-\frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|\right)}{\partial\sigma^2_{\boldsymbol{\alpha}}} &= -\frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'\right)+\frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'\right) \\
&= -\frac{1}{2}\text{trace}\left(\left(\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\right)\boldsymbol{Z}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'\right) \\
&= -\frac{1}{2}\text{trace}\left(\boldsymbol{P}\boldsymbol{Z}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'\right) \\
&= -\frac{1}{2}\text{trace}\left(\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\right),
\end{aligned}$$

Therefore,

$$\frac{\partial l}{\partial\sigma^2_{\boldsymbol{\alpha}}} = -\frac{1}{2}\text{trace}\left(\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\right)+\frac{1}{2}\hat{\boldsymbol{\alpha}}'\boldsymbol{G}^{-1}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{G}^{-1}\hat{\boldsymbol{\alpha}}.$$

If $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{U}_r\boldsymbol{\Sigma}^{-1/2}$ ($\boldsymbol{U}_r$ of dimension $c\times(c-q)$ contains the span or the non-null part of the singular value decomposition of $\boldsymbol{D}_q'\boldsymbol{D}_q$ and $\boldsymbol{\Sigma}$ a diagonal matrix of dimension $(c-q)\times(c-q)$ that contains the non-zero eigenvalues of the decomposition), $\boldsymbol{G} = \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{I}$

and $\frac{\partial \boldsymbol{G}}{\partial \sigma^2_{\boldsymbol{\alpha}}} = \boldsymbol{I}$, and we can write $\frac{\partial \boldsymbol{G}}{\partial \sigma^2_{\boldsymbol{\alpha}}} = \frac{1}{\sigma^4_{\boldsymbol{\alpha}}}\boldsymbol{GG}$. Therfore, replacing $\frac{\partial \boldsymbol{G}}{\sigma^2_{\boldsymbol{\alpha}}}$ by $\frac{1}{\sigma^4_{\boldsymbol{\alpha}}}\boldsymbol{GG}$ in $\frac{\partial l}{\partial \sigma^2_{\boldsymbol{\alpha}}}$:

$$2\frac{\partial l}{\partial \sigma^2_{\boldsymbol{\alpha}}} = -\frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\text{trace}\left(\boldsymbol{Z'PZ}\frac{\boldsymbol{GG}}{\sigma^2_{\boldsymbol{\alpha}}}\right) + \frac{1}{\sigma^4_{\boldsymbol{\alpha}}}\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}}.$$

Therefore:

$$\begin{aligned} 2\frac{\partial l}{\partial \sigma^2_{\boldsymbol{\alpha}}} = 0 &\Leftrightarrow \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\text{trace}\left(\boldsymbol{Z'PZ}\frac{\boldsymbol{GG}}{\sigma^2_{\boldsymbol{\alpha}}}\right) = \frac{1}{\sigma^4_{\boldsymbol{\alpha}}}\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}} \\ &\Leftrightarrow \sigma^2_{\boldsymbol{\alpha}} = \frac{\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}}}{\text{trace}\left(\boldsymbol{Z'PZ}\frac{\boldsymbol{GG}}{\sigma^2_{\boldsymbol{\alpha}}}\right)}. \end{aligned}$$

Estimation of ρ.
Derivative of Part I with respect to ρ:

$$\begin{aligned} \frac{\partial\left(\frac{1}{2}\log|\boldsymbol{V}|\right)}{\partial \sigma^2} &= \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \rho}\right) \\ &= \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{R}}{\partial \rho}\right). \end{aligned}$$

Derivative of Part II with respect to ρ:

$$\begin{aligned} \frac{\partial\left(\frac{1}{2}\log|\boldsymbol{X'V}^{-1}\boldsymbol{X}|\right)}{\partial \rho} &= \frac{1}{2}\text{trace}\left(\left(\boldsymbol{X'V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\frac{\partial \boldsymbol{V}^{-1}}{\partial \rho}\boldsymbol{X}\right) \\ &= -\frac{1}{2}\text{trace}\left(\left(\boldsymbol{X'V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X'V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \rho}\boldsymbol{V}^{-1}\boldsymbol{X}\right) \\ &= -\frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X'V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X'V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \rho}\right) \\ &= -\frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X'V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X'V}^{-1}\frac{\partial \boldsymbol{R}}{\partial \rho}\right). \end{aligned}$$

Derivative of Part III with respect to ρ:

$$\begin{aligned} \frac{\partial\left(\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\right)}{\partial \rho} &= \frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\frac{\partial \boldsymbol{V}^{-1}}{\partial \rho}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) \\ &= -\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \rho}\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) \\ &= -\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{R}}{\partial \rho}\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right). \end{aligned}$$

Joining the derivatives with respect to ρ of Part I and Part II:

$$\begin{aligned}\frac{\partial\left(-\frac{1}{2}\log|\boldsymbol{V}|-\frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|\right)}{\partial\rho} &= -\frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{R}}{\partial\rho}\right)+\frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{R}}{\partial\rho}\right)\\ &= -\frac{1}{2}\text{trace}\left(\left(\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\right)\frac{\partial\boldsymbol{R}}{\partial\rho}\right)\\ &= -\frac{1}{2}\text{trace}\left(\boldsymbol{P}\frac{\partial\boldsymbol{R}}{\partial\rho}\right),\end{aligned}$$

with $\boldsymbol{P}$ as in (A.1).

Therefore,

$$\frac{\partial l}{\partial\rho}=-\frac{1}{2}\text{trace}\left(\boldsymbol{P}\frac{\partial\boldsymbol{R}}{\partial\rho}\right)+\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{R}}{\partial\rho}\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right).$$

Second derivative of l with respect to ρ:

$$\begin{aligned}\frac{\partial^2 l}{\partial\rho^2} &= -\frac{1}{2}\text{trace}\left(\frac{\partial\boldsymbol{P}}{\partial\rho}\frac{\partial\boldsymbol{R}}{\partial\rho}+\boldsymbol{P}\frac{\partial^2\boldsymbol{R}}{\partial\rho^2}\right)+\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\frac{\partial\boldsymbol{V}^{-1}}{\partial\rho}\frac{\partial\boldsymbol{R}}{\partial\rho}\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\\ &+ \frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\frac{\partial^2\boldsymbol{R}}{\partial\rho^2}\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)+\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{R}}{\partial\rho}\frac{\partial\boldsymbol{V}^{-1}}{\partial\rho}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right),\end{aligned}$$

with $\boldsymbol{P}$ as in (A.1), i.e.

$$\begin{aligned}\frac{\partial\boldsymbol{P}}{\partial\rho} &= \frac{\partial\boldsymbol{V}^{-1}}{\partial\rho}-\frac{\partial\boldsymbol{V}^{-1}}{\partial\rho}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\\ &+ \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\frac{\partial\boldsymbol{V}^{-1}}{\partial\rho}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\\ &- \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\frac{\partial\boldsymbol{V}^{-1}}{\partial\rho},\end{aligned}$$

and $\frac{\partial\boldsymbol{V}^{-1}}{\partial\rho}=-\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\rho}\boldsymbol{V}^{-1}=-\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{R}}{\partial\rho}\boldsymbol{V}^{-1}$.

Notice that

$$\frac{\partial\boldsymbol{R}}{\partial\rho}=\sigma^2_{\boldsymbol{\epsilon}}\begin{bmatrix}0 & 1 & 2\rho & \cdots & (n-1)\rho^{n-2}\\ 1 & 0 & 1 & \cdots & (n-2)\rho^{n-3}\\ 2\rho & 1 & 0 & \cdots & (n-3)\rho^{n-4}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ (n-1)\rho^{n-2} & (n-2)\rho^{n-3} & (n-3)\rho^{n-4} & \cdots & 0\end{bmatrix},$$

which implies that

$$\frac{\partial^2 \boldsymbol{R}}{\partial \rho^2} = \sigma^2_{\boldsymbol{\epsilon}} \begin{bmatrix} 0 & 0 & 2 & \cdots & (n-1)(n-2)\rho^{n-3} \\ 0 & 0 & 0 & \cdots & (n-2)(n-3)\rho^{n-4} \\ 2 & 0 & 0 & \cdots & (n-3)(n-4)\rho^{n-5} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (n-1)(n-2)\rho^{n-3} & (n-2)(n-3)\rho^{n-4} & (n-3)(n-4)\rho^{n-5} & \cdots & 0 \end{bmatrix}.$$

To estimate the zero of $\frac{\partial l}{\partial \rho}$ we can:

- Use Newton-Raphson method:

$$\begin{aligned} \frac{\partial l}{\partial \rho} &\simeq \left.\frac{\partial l}{\partial \rho}\right|_{\rho_0} + \left.\frac{\partial^2 l}{\partial \rho^2}\right|_{\rho_0} (\rho_1 - \rho_0) \\ 0 &\simeq \left.\frac{\partial l}{\partial \rho}\right|_{\rho_0} + \left.\frac{\partial^2 l}{\partial \rho^2}\right|_{\rho_0} (\rho_1 - \rho_0) \\ \left.\frac{\partial^2 l}{\partial \rho^2}\right|_{\rho_0} \rho_1 &\simeq \left.\frac{\partial^2 l}{\partial \rho^2}\right|_{\rho_0} \rho_0 - \left.\frac{\partial l}{\partial \rho}\right|_{\rho_0} \rho_1 \\ \rho_1 &\simeq \rho_0 - \frac{\left.\frac{\partial^2 l}{\partial \rho^2}\right|_{\rho_0}}{\left.\frac{\partial l}{\partial \rho}\right|_{\rho_0}}. \end{aligned}$$

- Use the nleqslv function of the R package also called nleqslv.

Instead of solving the non linear equation we can use the function optim of the R package stats to get ρ that minimizes the approximate restricted maximum likelihood for some given values of $\sigma^2_{\boldsymbol{\epsilon}}$ and $\sigma^2_{\boldsymbol{\alpha}}$.

A.2.1 R code to estimate the variance and correlation parameters

```
CovMatriz = function(n, sig2, rho){
    # Covariance matrix
    times = 1:n
    H = abs(outer(times, times, "-"))
    R = sig2 * rho^H
    R
}

SchallRho = function(y, X, Z) {
  # Input
  # data: y
  # model matrices: X, Z
```

```
init.time = proc.time()[3]

k = dim(X)[2]; q = dim(Z)[2]; n = length(y)

itermax = 400; tol = 1e-05

sig2.r.new = 0.1; sig2.new = 0.1; rho.new = 0
iter = 0
dif = tol + 1
while (iter < itermax & dif >= tol) {

  iter = iter + 1
  sig2.r = sig2.r.new
  sig2 = sig2.new
  rho = rho.new

  R = CovMatriz(n, sig2, rho); R.inv = solve(R)
  G = sig2.r * diag(ncol(Z)); G.inv = solve(G)
  V = R + Z %*% G %*% t(Z)
  V.inv = R.inv - R.inv %*% Z %*% solve(G.inv + t(Z) %*% R.inv %*% Z) %*% t(Z) %*% R.inv
  P = V.inv - V.inv %*% X %*% solve(t(X) %*% V.inv %*% X) %*% t(X) %*% V.inv
  Rtilde = R * (1/sig2); Rtilde.inv = sig2 * R.inv

  beta = solve(t(X) %*% V.inv %*% X) %*% t(X) %*% V.inv %*% y
  alpha = G %*% t(Z) %*% V.inv %*% (y- X %*% beta)

  num = t(alpha) %*% alpha
  den = tr(t(Z) %*% P %*% Z %*% (G %*% G)/sig2.r)
  sig2.r.new = as.numeric(num/den)

  a = (y - X %*% beta - Z %*% alpha)
  sig2.new =  as.numeric((t(a) %*% Rtilde.inv %*% a)/tr(P %*% R))

  REML1D = function(rho){
    R = CovMatriz(n, sig2, rho); R.inv = solve(R)
    V = R +  Z %*% G %*% t(Z);
    V.inv = R.inv - R.inv %*% Z %*% solve(G.inv + t(Z) %*% R.inv %*% Z) %*% t(Z) %*% R.inv
    P = V.inv - V.inv %*% X %*% solve(t(X) %*% V.inv %*% X) %*% t(X) %*% V.inv

    log.like = (1/2) * sum(log(eigen(V)$values)) + (1/2) * sum(log(eigen(t(X) %*%
               V.inv %*% X)$values)) + (1/2) * t(y - X %*% beta) %*% V.inv %*% (y - X %*% beta)
    return(log.like)
  }
  optimREML = optim(rho, REML1D, NULL, method = "L-BFGS-B", lower = -0.99999999,
   upper = 0.999999999)
```

```
    Valorlog.like = optimREML$value
    rho.new = as.numeric(optimREML$par)


    par.new = c(sig2.new, sig2.r.new, rho.new)
    par = c(sig2, sig2.r, rho)

    h = 0
    for (i in 1:3) {
      h = h + abs(par.new[i] - par[i])
    }
    dif = h/3
  }
  end.time = proc.time()[3]
  schall = list(beta, alpha, sig2.r.new, sig2.new, rho.new, iter, end.time - init.time)
}
```

Appendix B

Appendix to Chapter 3

In this appendix we include the proofs of Corollary 3.1 and of Theorem 3.2.

B.1 Proof of Corollary 3.1

Proof. Notice that if $\boldsymbol{P}^{\boldsymbol{z}}_{+22} = \boldsymbol{O}$, by (3.19), the coefficients that give the fit are:

$$\hat{\boldsymbol{\theta}}_{+_{1,\ldots,c}} = \left(\boldsymbol{B}'\boldsymbol{B} + \lambda_{\boldsymbol{x}}(\boldsymbol{D}'_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{x}} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}}) + \lambda_{\boldsymbol{z}}(\boldsymbol{I}_{c_{\boldsymbol{x}}} \otimes \boldsymbol{D}'_{\boldsymbol{z}}\boldsymbol{D}_{\boldsymbol{z}})\right)^{-1}\boldsymbol{B}'\boldsymbol{y},$$

i.e., the same as the coefficients we obtain only fitting the data without a prediction, (3.7). Let us see which are the coefficients that determine the forecast when the penalty orders are two or three.

- Differences of order 2.

  Suppose a difference matrix with second order penalty $\boldsymbol{D}_{\boldsymbol{x}_+}$ of dimensions $(c_{\boldsymbol{x}_+} - 2) \times c_{\boldsymbol{x}_+}$:

$$\boldsymbol{D}_{\boldsymbol{x}_+} = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{O} \\ \boldsymbol{D}_{\boldsymbol{x}(1)} & \boldsymbol{D}_{\boldsymbol{x}(2)} \end{bmatrix} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & -2 & 1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix},$$

  where $\boldsymbol{D}_{\boldsymbol{x}(2)}$:

$$\boldsymbol{D}_{\boldsymbol{x}(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ -2 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix},$$

of dimension $c_{\boldsymbol{x}_p} \times c_{\boldsymbol{x}_p}$, where $c_{\boldsymbol{x}_p}$ is the number of columns of $\boldsymbol{B}_{\boldsymbol{x}(2)}$. Therefore, $(\boldsymbol{D}_{\boldsymbol{x}(2)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}})^{-1}$ has the form

$$\begin{bmatrix}
1 & & & & & & & & & & & \\
 & \ddots & & & & & & & & & & \\
 & & 1 & & & & & & & & & \\
2 & & & 1 & & & & & & & & \\
 & \ddots & & & \ddots & & & & & & & \\
 & & 2 & & & 1 & & & & & & \\
3 & & & 2 & & & 1 & & & & & \\
 & \ddots & & & \ddots & & & \ddots & & & & \\
 & & 3 & & & 2 & & & 1 & & & \\
4 & & & 3 & & & 2 & & & 1 & & \\
 & \ddots & & & \ddots & & & \ddots & & & \ddots & \\
 & & 4 & & & 3 & & & 2 & & & 1 \\
 & \vdots & & & \vdots & & & \vdots & & & \vdots & & \ddots
\end{bmatrix}$$

of dimension $c_{\boldsymbol{x}_p} c_{\boldsymbol{z}} \times c_{\boldsymbol{x}_p} c_{\boldsymbol{z}}$, each block has dimension $c_{\boldsymbol{z}} \times c_{\boldsymbol{z}}$. Moreover,

$$\boldsymbol{D}_{\boldsymbol{x}(1)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}} = \begin{bmatrix}
0 & & 0 & & & & 1 & & -2 & & \\
 & \ddots & & \ddots & & \cdots & & \ddots & & & \ddots & \\
 & & 0 & & 0 & & & & 1 & & & -2 \\
0 & & & 0 & & & 0 & & & 1 & & \\
 & \ddots & & & \ddots & \cdots & & \ddots & & & \ddots & \\
 & & 0 & & & 0 & & & 0 & & & 1 \\
0 & & & 0 & & & 0 & & & 0 & & \\
 & \ddots & & & \ddots & \cdots & & \ddots & & & \ddots & \\
 & & 0 & & & 0 & & & 0 & & & 0 \\
 & \vdots & & & \vdots & \cdots & & \vdots & & & \vdots & &
\end{bmatrix},$$

i.e., is a matrix of dimension $c_{\boldsymbol{x}_p} c_{\boldsymbol{z}} \times c_{\boldsymbol{x}} c_{\boldsymbol{z}}$ with just three blocks of dimension

$c_{\boldsymbol{z}} \times c_{\boldsymbol{z}}$ that are not blocks of zeros. Therefore,

$$(\boldsymbol{D}_{\boldsymbol{x}(2)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}})^{-1}(\boldsymbol{D}_{\boldsymbol{x}(1)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}}) = \begin{bmatrix} 0 & & 0 & & & 1 & & -2 & & \\ & \ddots & & \ddots & \cdots & & \ddots & & \ddots & \\ & 0 & & 0 & & & 1 & & -2 \\ 0 & & 0 & & & 2 & & -3 & & \\ & \ddots & & \ddots & \cdots & & \ddots & & \ddots & \\ & 0 & & 0 & & & 2 & & -3 \\ 0 & & 0 & & & 3 & & -4 & & \\ & \ddots & & \ddots & \cdots & & \ddots & & \ddots & \\ & 0 & & 0 & & & 3 & & -4 \\ 0 & & 0 & & & 4 & & -5 & & \\ & \ddots & & \ddots & \cdots & & \ddots & & \ddots & \\ & 0 & & 0 & & & 4 & & -5 \\ & \vdots & & \vdots & \cdots & & \vdots & & \vdots & \end{bmatrix},$$

with dimension $c_{\boldsymbol{x}_p} c_{\boldsymbol{z}} \times c_{\boldsymbol{x}} c_{\boldsymbol{z}}$. Hence, considering the matrix of coefficients that give the fit, $\hat{\boldsymbol{\Theta}}$, and the matrix of coefficients that give the forecast, $\hat{\boldsymbol{\Theta}}_p$, each row $j = 1, ..., c_{\boldsymbol{z}}$, of the additional matrix of coefficients is a linear combination of two old coefficients of that row:

$$\hat{\boldsymbol{\Theta}}_{j\ .} = \hat{\theta}_{j\ c_{\boldsymbol{x}}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} + (\hat{\theta}_{j\ c_{\boldsymbol{x}}} - \hat{\theta}_{j\ c_{\boldsymbol{x}}-1}) \begin{bmatrix} 1 \\ 2 \\ 3 \\ \vdots \end{bmatrix}.$$

- Differences of order 3.
  Suppose a difference matrix with third order penalty, $\boldsymbol{D}_{\boldsymbol{x}_+}$ of dimensions $(c_{\boldsymbol{x}_+} - 3) \times c_{\boldsymbol{x}_+}$:

$$\boldsymbol{D}_+ = \begin{bmatrix} -1 & 3 & -3 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & -1 & 3 & -3 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 3 & -3 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & -1 & 3 & -3 & 1 \end{bmatrix}.$$

  In this case, $\boldsymbol{D}_{\boldsymbol{x}(2)}$ is:

$$\boldsymbol{D}_{\boldsymbol{x}(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -3 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 3 & -3 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 3 & -3 & 1 \end{bmatrix}.$$

  $(\boldsymbol{D}_{\boldsymbol{x}(2)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}})^{-1}$ and $\boldsymbol{D}_{\boldsymbol{x}(1)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}}$ are:

$$(\boldsymbol{D}_{\boldsymbol{x}(2)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}})^{-1} = \begin{bmatrix} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ 3 & & & 1 & & & & & & & \\ & \ddots & & & \ddots & & & & & & \\ & & 3 & & & 1 & & & & & \\ 6 & & & 3 & & & 1 & & & & \\ & \ddots & & & \ddots & & & \ddots & & & \\ & & 6 & & & 3 & & & 1 & & \\ 10 & & & 6 & & & 3 & & & 1 & \\ & \ddots & & & \ddots & & & \ddots & & & \ddots \\ & & 10 & & & 6 & & & 3 & & \\ & \vdots & & & \vdots & & & \vdots & & & \end{bmatrix},$$

$$\boldsymbol{D}_{\boldsymbol{x}(1)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}} = \begin{bmatrix} 0 & & & -1 & & 3 & & -3 & & \\ & \ddots & \dots & & \ddots & & \ddots & & \ddots & \\ & 0 & & & -1 & & 3 & & -3 & \\ 0 & & & 0 & & -1 & & 3 & & \\ & \ddots & \dots & & \ddots & & \ddots & & \ddots & \\ & 0 & & & 0 & & -1 & & 3 & \\ 0 & & & 0 & & 0 & & -1 & & \\ & \ddots & \dots & & \ddots & & \ddots & & \ddots & \\ & 0 & & & 0 & & 0 & & -1 & \\ & \vdots & \dots & & \vdots & & \vdots & & \vdots & \end{bmatrix}.$$

Then,

$$(\boldsymbol{D}_{\boldsymbol{x}(2)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}})^{-1}(\boldsymbol{D}_{\boldsymbol{x}(1)} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}}}) = \begin{bmatrix} 0 & & & -1 & & 3 & & -3 & & \\ & \ddots & \dots & & \ddots & & \ddots & & \ddots & \\ & 0 & & & -1 & & 3 & & & -3 \\ 0 & & & -3 & & 8 & & -6 & & \\ & \ddots & \dots & & \ddots & & \ddots & & \ddots & \\ & 0 & & & -3 & & 8 & & & -6 \\ 0 & & & -6 & & 15 & & -10 & & \\ & \ddots & \dots & & \ddots & & \ddots & & \ddots & \\ & 0 & & & -6 & & 15 & & & -10 \\ 0 & & & -10 & & 24 & & -15 & & \\ & \ddots & \dots & & \ddots & & \ddots & & \ddots & \\ & 0 & & & -10 & & 24 & & & -15 \\ & \vdots & \dots & & \vdots & & \vdots & & \vdots & \end{bmatrix}.$$

Therefore, each row, $j = 1, ..., c_{\boldsymbol{z}}$, of the additional matrix of coefficients is a linear

combination of three old coefficients of that row:

$$\hat{\mathbf{\Theta}}_{p_j\ .} = \hat{\theta}_{j\ c_{\boldsymbol{x}}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} + \frac{3\hat{\theta}_{j\ c_{\boldsymbol{x}}} - 4\hat{\theta}_{j\ c_{\boldsymbol{x}}-1} + \hat{\theta}_{j\ c_{\boldsymbol{x}}-2}}{2} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ \vdots \end{bmatrix} + \frac{\hat{\theta}_{j\ c_{\boldsymbol{x}}} - 2\hat{\theta}_{j\ c_{\boldsymbol{x}}-1} + \hat{\theta}_{j\ c_{\boldsymbol{x}}-2}}{2} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ \vdots \end{bmatrix}^2 .$$

■

B.2 Proof of Theorem 3.2

Proof. Given the extended transformation $\boldsymbol{\Omega}_+^*$ in two dimensions defined in (3.39) and the extended penalty matrix in (3.12) the extended mixed model penalty is:

$$
\begin{aligned}
\boldsymbol{\Phi}_+^* &= \boldsymbol{\Omega}_+^{*'}\boldsymbol{P}_+\boldsymbol{\Omega}_+^* = \boldsymbol{\Omega}_+^{*'}(\lambda_{\boldsymbol{z}}\boldsymbol{P}_+^{\boldsymbol{z}+} + \lambda_{\boldsymbol{x}}\boldsymbol{P}_+^{\boldsymbol{x}+})\boldsymbol{\Omega}_+^* = \lambda_{\boldsymbol{z}}\boldsymbol{\Omega}_+^{*'}\boldsymbol{P}_+^{\boldsymbol{z}+}\boldsymbol{\Omega}_+^* + \lambda_{\boldsymbol{x}}\boldsymbol{\Omega}_+^{*'}\boldsymbol{P}_+^{\boldsymbol{x}+}\boldsymbol{\Omega}_+^* \\
&= \lambda_{\boldsymbol{z}}\begin{bmatrix} \boldsymbol{U}_{\boldsymbol{x}+f}^{*'}\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^{*'} \\ \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^{*'} \\ \boldsymbol{U}_{\boldsymbol{x}+f}^{*'}\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'} \\ \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'} \end{bmatrix}(\boldsymbol{I}_{c_{\boldsymbol{x}+}}\otimes\boldsymbol{D}_{\boldsymbol{z}+}'\boldsymbol{D}_{\boldsymbol{z}+})\left[\boldsymbol{U}_{\boldsymbol{x}+f}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^* \mid \boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^* \mid \boldsymbol{U}_{\boldsymbol{x}+f}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^* \mid \boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^*\right] \\
&+ \lambda_{\boldsymbol{x}}\begin{bmatrix} \boldsymbol{U}_{\boldsymbol{x}+f}^{*'}\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^{*'} \\ \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^{*'} \\ \boldsymbol{U}_{\boldsymbol{x}+f}^{*'}\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'} \\ \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'} \end{bmatrix}(\boldsymbol{D}_{\boldsymbol{x}+}'\boldsymbol{D}_{\boldsymbol{x}+}\otimes\boldsymbol{I}_{c_{\boldsymbol{z}+}})\left[\boldsymbol{U}_{\boldsymbol{x}+f}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^* \mid \boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^* \mid \boldsymbol{U}_{\boldsymbol{x}+f}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^* \mid \boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^*\right] \\
&= \lambda_{\boldsymbol{z}}\begin{bmatrix} \boldsymbol{U}_{\boldsymbol{x}+f}^{*'}\boldsymbol{I}_{c_{\boldsymbol{x}+}}\boldsymbol{U}_{\boldsymbol{x}+f}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^{*'}\boldsymbol{D}_{\boldsymbol{z}}'\boldsymbol{D}_{\boldsymbol{z}+}\boldsymbol{U}_{\boldsymbol{z}+f}^* & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{I}_{c_{\boldsymbol{x}+}}\boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^{*'}\boldsymbol{D}_{\boldsymbol{z}+}'\boldsymbol{D}_{\boldsymbol{z}+}\boldsymbol{U}_{\boldsymbol{z}+f}^* \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix} \\
&\quad \begin{bmatrix} \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{U}_{\boldsymbol{x}+f}^{*'}\boldsymbol{I}_{c_{\boldsymbol{x}+}}\boldsymbol{U}_{\boldsymbol{x}+f}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'}\boldsymbol{D}_{\boldsymbol{z}+}'\boldsymbol{D}_{\boldsymbol{z}+}\boldsymbol{U}_{\boldsymbol{z}+r}^* & \boldsymbol{U}_{\boldsymbol{x}+f}^{*'}\boldsymbol{I}_{c_{\boldsymbol{x}+}}\boldsymbol{U}_{\boldsymbol{x}+f}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'}\boldsymbol{D}_{\boldsymbol{z}+}'\boldsymbol{D}_{\boldsymbol{z}+}\boldsymbol{U}_{\boldsymbol{z}+r}^* \\ \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{I}_{c_{\boldsymbol{x}+}}\boldsymbol{U}_{\boldsymbol{x}+f}^*\otimes\boldsymbol{U}_{\boldsymbol{z}r}^{*'}\boldsymbol{D}_{\boldsymbol{z}+}'\boldsymbol{D}_{\boldsymbol{z}+}\boldsymbol{U}_{\boldsymbol{z}r}^* & \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{I}_{c_{\boldsymbol{x}+}}\boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'}\boldsymbol{D}_{\boldsymbol{z}+}'\boldsymbol{D}_{\boldsymbol{z}+}\boldsymbol{U}_{\boldsymbol{z}+r}^* \end{bmatrix} \\
&+ \lambda_{\boldsymbol{x}}\begin{bmatrix} \boldsymbol{U}_{\boldsymbol{x}+f}^{*'}\boldsymbol{D}_{\boldsymbol{x}+}'\boldsymbol{D}_{\boldsymbol{x}+}\boldsymbol{U}_{\boldsymbol{x}+f}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^{*'}\boldsymbol{U}_{\boldsymbol{z}+f}^* & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{D}_{\boldsymbol{x}+}'\boldsymbol{D}_{\boldsymbol{x}+}\boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^{*'}\boldsymbol{U}_{\boldsymbol{z}+f}^* \\ \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{D}_{\boldsymbol{x}+}'\boldsymbol{D}_{\boldsymbol{x}+}\boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'}\boldsymbol{U}_{\boldsymbol{z}+f}^* & \boldsymbol{O} \end{bmatrix} \\
&\quad \begin{bmatrix} \boldsymbol{O} & \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{D}_{\boldsymbol{x}+}'\boldsymbol{D}_{\boldsymbol{x}+}\boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^{*'}\boldsymbol{U}_{\boldsymbol{z}+r}^* \\ \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{U}_{\boldsymbol{x}+f}^{*'}\boldsymbol{D}_{\boldsymbol{x}+}'\boldsymbol{D}_{\boldsymbol{x}+}\boldsymbol{U}_{\boldsymbol{x}+f}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'}\boldsymbol{U}_{\boldsymbol{z}+r}^* & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{D}_{\boldsymbol{x}+}'\boldsymbol{D}_{\boldsymbol{x}+}\boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'}\boldsymbol{U}_{\boldsymbol{z}+r}^* \end{bmatrix} \\
&= \lambda_{\boldsymbol{z}}\begin{bmatrix} \boldsymbol{U}_{\boldsymbol{x}+f}^{*'}\boldsymbol{I}_{c_{\boldsymbol{x}+}}\boldsymbol{U}_{\boldsymbol{x}+f}^*\otimes\boldsymbol{O}_{q_{\boldsymbol{z}}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{I}_{c_{\boldsymbol{x}+}}\boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{O}_{q_{\boldsymbol{z}}} \\ \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix} \\
&\quad \begin{bmatrix} \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{U}_{\boldsymbol{x}+f}^{*'}\boldsymbol{U}_{\boldsymbol{x}+f}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'}\boldsymbol{D}_{\boldsymbol{z}+}'\boldsymbol{D}_{\boldsymbol{z}+}\boldsymbol{U}_{\boldsymbol{z}+r}^* & \boldsymbol{U}_{\boldsymbol{x}+f}^{*'}\boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'}\boldsymbol{D}_{\boldsymbol{z}+}'\boldsymbol{D}_{\boldsymbol{z}+}\boldsymbol{U}_{\boldsymbol{z}+r}^* \\ \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{U}_{\boldsymbol{x}+f}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'}\boldsymbol{D}_{\boldsymbol{z}+}'\boldsymbol{D}_{\boldsymbol{z}+}\boldsymbol{U}_{\boldsymbol{z}+r}^* & \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'}\boldsymbol{D}_{\boldsymbol{z}+}'\boldsymbol{D}_{\boldsymbol{z}+}\boldsymbol{U}_{\boldsymbol{z}+r}^* \end{bmatrix} \\
&+ \lambda_{\boldsymbol{x}}\begin{bmatrix} \boldsymbol{O}_{q_{\boldsymbol{x}}}\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^{*'}\boldsymbol{I}_{c_{\boldsymbol{z}+}}\boldsymbol{U}_{\boldsymbol{z}+f}^{*'} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{D}_{\boldsymbol{x}+}'\boldsymbol{D}_{\boldsymbol{x}+}\boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^{*'}\boldsymbol{U}_{\boldsymbol{z}+f}^* \\ \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{D}_{\boldsymbol{x}+}'\boldsymbol{D}_{\boldsymbol{x}+}\boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'}\boldsymbol{U}_{\boldsymbol{z}+f}^* \end{bmatrix} \\
&\quad \begin{bmatrix} \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{D}_{\boldsymbol{x}+}'\boldsymbol{D}_{\boldsymbol{x}+}\boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+f}^{*'}\boldsymbol{U}_{\boldsymbol{z}+r}^* \\ \boldsymbol{O}_{q_{\boldsymbol{x}}}\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'}\boldsymbol{U}_{\boldsymbol{z}+r}^* & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{U}_{\boldsymbol{x}+r}^{*'}\boldsymbol{D}_{\boldsymbol{x}+}'\boldsymbol{D}_{\boldsymbol{x}+}\boldsymbol{U}_{\boldsymbol{x}+r}^*\otimes\boldsymbol{U}_{\boldsymbol{z}+r}^{*'}\boldsymbol{U}_{\boldsymbol{z}+r}^* \end{bmatrix}
\end{aligned}
$$

Therefore, $\boldsymbol{\Phi}_+^* = \text{blockdiag}(\boldsymbol{O}_{q_{\boldsymbol{z}}q_{\boldsymbol{x}}}, \boldsymbol{F}_+^*)$, with $\boldsymbol{F}_+^*$ given in (3.42). Then, the extended covariance matrix of the random effects is $\boldsymbol{G}_+^* = \sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{F}_+^{*-1}$. ■

Appendix C

Appendix to Chapter 4

C.1 Proof of Theorem 4.1

Proof. Given the extended transformation matrix for the random part $\boldsymbol{\Omega}_{+r}$ and the extended penalty matrix $\boldsymbol{P}_+$ defined in (4.8) and (4.7), respectively, $\boldsymbol{F}_+$ is:

$$
\begin{aligned}
\boldsymbol{F}_+ &= \boldsymbol{\Omega}'_{r+}\boldsymbol{P}_+\boldsymbol{\Omega}_{r+} \\
&= \begin{bmatrix} 0 & \boldsymbol{U}'_{\boldsymbol{z}r} & \cdots & & \\ \vdots & & \boldsymbol{U}'_{\boldsymbol{x}_+r} & & \\ & & & \boldsymbol{u}^{(2)\prime}_{\boldsymbol{x}_+f} \otimes \boldsymbol{U}'_{\boldsymbol{z}_+r} & \\ & & & \boldsymbol{U}'_{\boldsymbol{x}_+r} \otimes \boldsymbol{u}^{(2)\prime}_{\boldsymbol{z}_+f} & \\ & & & \boldsymbol{U}'_{\boldsymbol{x}_+r} \otimes \boldsymbol{U}'_{\boldsymbol{z}_+r} & \end{bmatrix} \begin{bmatrix} 0 & \cdots & & \\ \vdots & \lambda_{\boldsymbol{z}}\boldsymbol{D}'_{\boldsymbol{z}_+}\boldsymbol{D}_{\boldsymbol{z}_+} & & \\ & & \lambda_{\boldsymbol{x}}\boldsymbol{D}'_{\boldsymbol{x}_+}\boldsymbol{D}_{\boldsymbol{x}_+} & \\ & & & \tau_{\boldsymbol{x}}\boldsymbol{D}'_{\boldsymbol{x}_+}\boldsymbol{D}_{\boldsymbol{x}_+} \otimes \boldsymbol{I}_{c_{\boldsymbol{z}_+}} + \tau_{\boldsymbol{z}}\boldsymbol{I}_{c_{\boldsymbol{x}_+}} \otimes \boldsymbol{D}'_{\boldsymbol{z}_+}\boldsymbol{D}_{\boldsymbol{z}_+} \end{bmatrix} \\
&\quad \begin{bmatrix} 0 & \cdots & & \\ \boldsymbol{U}_{\boldsymbol{z}r} & & & \\ \vdots & \boldsymbol{U}_{\boldsymbol{x}_+r} & & \\ & \boldsymbol{u}^{(2)}_{\boldsymbol{x}_+f} \otimes \boldsymbol{U}_{\boldsymbol{z}_+r} \mid \boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{u}^{(2)}_{\boldsymbol{z}_+f} \mid \boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{U}_{\boldsymbol{z}_+r} & \end{bmatrix} \\
&= \begin{bmatrix} \lambda_{\boldsymbol{z}}\boldsymbol{U}'_{\boldsymbol{z}_+f}\boldsymbol{D}'_{\boldsymbol{z}_+}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+f} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \lambda_{\boldsymbol{x}}\boldsymbol{U}'_{\boldsymbol{x}_+f}\boldsymbol{D}'_{\boldsymbol{x}_+}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+f} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \tau_{\boldsymbol{x}}\boldsymbol{u}^{(2)\prime}_{\boldsymbol{x}_+f}\boldsymbol{D}'_{\boldsymbol{x}_+}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{u}^{(2)}_{\boldsymbol{x}_+} \otimes \boldsymbol{U}'_{\boldsymbol{z}_+r}\boldsymbol{U}_{\boldsymbol{z}_+r} + \tau_{\boldsymbol{z}}\boldsymbol{u}^{(2)\prime}_{\boldsymbol{x}_+f}\boldsymbol{u}^{(2)}_{\boldsymbol{x}_+f} \otimes \boldsymbol{U}'_{\boldsymbol{z}_+r}\boldsymbol{D}'_{\boldsymbol{z}_+}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \tau_{\boldsymbol{x}}\boldsymbol{U}'_{\boldsymbol{x}_+r}\boldsymbol{D}'_{\boldsymbol{x}_+}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{u}^{(2)\prime}_{\boldsymbol{z}_+f}\boldsymbol{u}^{(2)}_{\boldsymbol{z}_+f} + \tau_{\boldsymbol{z}}\boldsymbol{U}'_{\boldsymbol{x}_+r}\boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{u}^{(2)\prime}_{\boldsymbol{z}_+f}\boldsymbol{D}'_{\boldsymbol{z}_+}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{u}^{(2)}_{\boldsymbol{z}_+f} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \tau_{\boldsymbol{x}}\boldsymbol{U}'_{\boldsymbol{x}_+r}\boldsymbol{D}'_{\boldsymbol{x}_+}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{U}'_{\boldsymbol{z}_+r}\boldsymbol{U}_{\boldsymbol{z}_+r} + \tau_{\boldsymbol{z}}\boldsymbol{U}'_{\boldsymbol{x}_+r}\boldsymbol{U}_{\boldsymbol{x}_+r} \otimes \boldsymbol{U}'_{\boldsymbol{z}_+r}\boldsymbol{D}'_{\boldsymbol{z}_+}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r} \end{bmatrix},
\end{aligned}
$$

using $\boldsymbol{U}'_{ir}\boldsymbol{D}'_i\boldsymbol{D}_i\boldsymbol{U}_{ir} = \tilde{\boldsymbol{\Sigma}}_i$, $\boldsymbol{u}^{(2)\prime}_{if}\boldsymbol{D}'_i = \boldsymbol{O}$, $\boldsymbol{u}^{(2)\prime}_{if}\boldsymbol{u}^{(2)}_{if} = 1$ and $\boldsymbol{U}'_{ir}\boldsymbol{U}_{ir} = \boldsymbol{I}_{c_i-q_i}$, for $i = \boldsymbol{z}_+, \boldsymbol{x}_+$, we obtain the extended mixed model penalty $\boldsymbol{F}_+$ in (4.11). ∎

C.2 Proof of Theorem 4.2

Proof. Given the extended transformation matrix for the random part $\boldsymbol{\Omega}_{+r}$ and the extended penalty matrix $\boldsymbol{P}_+$ defined in (4.15) and (4.7), respectively, $\boldsymbol{F}_+^*$ is:

$$\boldsymbol{F}_+^* = \boldsymbol{\Omega}_{r+}^{*\prime}\boldsymbol{P}_+\boldsymbol{\Omega}_{r+}^* = \begin{bmatrix} 0 & \boldsymbol{U}_{\boldsymbol{z}_+r}^{*\prime} & \cdots & \\ \vdots & & & \\ & \boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime} & & \\ & & \boldsymbol{u}_{\boldsymbol{x}_+f}^{*(2)\prime}\otimes\boldsymbol{U}_{\boldsymbol{z}_+r}^{*\prime} & \\ & & \boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\otimes\boldsymbol{u}_{\boldsymbol{z}_+f}^{*(2)\prime} & \\ & & \boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\otimes\boldsymbol{U}_{\boldsymbol{z}_+r}^{*\prime} & \end{bmatrix} \begin{bmatrix} 0 & \cdots & \\ \vdots & & \\ \lambda_{\boldsymbol{z}}\boldsymbol{D}_{\boldsymbol{z}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{z}_+} & & \\ & \lambda_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{x}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_+} & \\ & & \lambda_3\boldsymbol{D}_{\boldsymbol{x}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_+}\otimes\boldsymbol{I}_{c_{\boldsymbol{z}_+}}+\lambda_4\boldsymbol{I}_{c_{\boldsymbol{x}_+}}\otimes\boldsymbol{D}_{\boldsymbol{z}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{z}_+} \end{bmatrix}$$

$$\begin{bmatrix} 0 & \cdots & \\ \boldsymbol{U}_{\boldsymbol{z}r} & & \\ \vdots & \boldsymbol{U}_{\boldsymbol{x}_+r}^{*} & \\ & \boldsymbol{u}_{\boldsymbol{x}_+f}^{*(2)}\otimes\boldsymbol{U}_{\boldsymbol{z}_+r}^{*} \mid \boldsymbol{U}_{\boldsymbol{x}_+r}^{*}\otimes\boldsymbol{u}_{\boldsymbol{z}_+f}^{*(2)} \mid \boldsymbol{U}_{\boldsymbol{x}_+r}^{*}\otimes\boldsymbol{U}_{\boldsymbol{z}_+r}^{*} & \end{bmatrix}$$

$$= \begin{bmatrix} \lambda_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{z}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{z}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r}^{*} & & \\ & \lambda_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{x}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*} & \\ & & \boldsymbol{F}_+^{(1,2)} \end{bmatrix},$$

where $\boldsymbol{F}_+^{(1,2)} = \begin{bmatrix} \boldsymbol{F}_{+11}^{(1,2)} & \boldsymbol{O} & \boldsymbol{F}_{+13}^{(1,2)} \\ \boldsymbol{O} & \boldsymbol{F}_{+22}^{(1,2)} & \boldsymbol{O} \\ \boldsymbol{F}_{+13}^{(1,2)\prime} & \boldsymbol{O} & \boldsymbol{F}_{+33}^{(1,2)} \end{bmatrix}$, with

$$\begin{aligned}
\boldsymbol{F}_{+11}^{(1,2)} &= \tau_{\boldsymbol{x}}\boldsymbol{u}_{\boldsymbol{x}_+f}^{*(2)\prime}\boldsymbol{D}_{\boldsymbol{x}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{u}_{\boldsymbol{x}_+f}^{*(2)}\otimes\boldsymbol{U}_{\boldsymbol{z}r}^{*\prime}\boldsymbol{U}_{\boldsymbol{z}_+r}^{*} + \tau_{\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{x}_+f}^{*(2)\prime}\boldsymbol{u}_{\boldsymbol{x}_+f}^{*(2)}\otimes\boldsymbol{U}_{\boldsymbol{z}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{z}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r}^{*} \\
\boldsymbol{F}_{+13}^{(1,2)} &= \tau_{\boldsymbol{z}}\boldsymbol{u}_{\boldsymbol{x}_+f}^{*(2)\prime}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*}\otimes\boldsymbol{U}_{\boldsymbol{z}_+r}^{*}\boldsymbol{D}_{\boldsymbol{z}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r}^{*} + \tau_{\boldsymbol{x}}\boldsymbol{u}_{\boldsymbol{x}_+f}^{*(2)\prime}\boldsymbol{D}_{\boldsymbol{x}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*}\otimes\boldsymbol{U}_{\boldsymbol{z}_+r}^{*\prime}\boldsymbol{U}_{\boldsymbol{z}_+r}^{*} \\
\boldsymbol{F}_{+22}^{(1,2)} &= \tau_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{x}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*}\otimes\boldsymbol{u}_{\boldsymbol{z}_+f}^{(*2)\prime}\boldsymbol{u}_{\boldsymbol{z}_+f}^{*(2)} + \tau_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*}\otimes\boldsymbol{u}_{\boldsymbol{z}_+f}^{*(2)\prime}\boldsymbol{D}_{\boldsymbol{z}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{u}_{\boldsymbol{z}_+f}^{*(2)} \\
\boldsymbol{F}_{+23}^{(1,2)} &= \tau_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{x}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*}\otimes\boldsymbol{u}_{\boldsymbol{z}_+f}^{*(2)\prime}\boldsymbol{U}_{\boldsymbol{z}_+r}^{*} + \tau_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*}\otimes\boldsymbol{u}_{\boldsymbol{z}_+f}^{*(2)\prime}\boldsymbol{D}_{\boldsymbol{z}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r}^{*} \\
\boldsymbol{F}_{+33}^{(1,2)} &= \tau_{\boldsymbol{z}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*}\otimes\boldsymbol{U}_{\boldsymbol{z}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{z}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{z}_+}\boldsymbol{U}_{\boldsymbol{z}_+r}^{*} + \tau_{\boldsymbol{x}}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*\prime}\boldsymbol{D}_{\boldsymbol{x}_+}^{\prime}\boldsymbol{D}_{\boldsymbol{x}_+}\boldsymbol{U}_{\boldsymbol{x}_+r}^{*}\otimes\boldsymbol{U}_{\boldsymbol{z}_+r}^{*\prime}\boldsymbol{U}_{\boldsymbol{z}_+r}^{*}
\end{aligned}$$

using $\boldsymbol{u}_{if}^{*(2)\prime}\boldsymbol{D}_i^{\prime} = \boldsymbol{O}$ for $i = \boldsymbol{z}_+, \boldsymbol{x}_+$, we obtain the extended mixed model penalty $\boldsymbol{F}_+^*$ in (4.18). ■

C.3 Simulation study results

Simulations results for Scenario 1:

- $n_{\boldsymbol{z}_p} = 0$, $n_{\boldsymbol{x}_p} = 5$

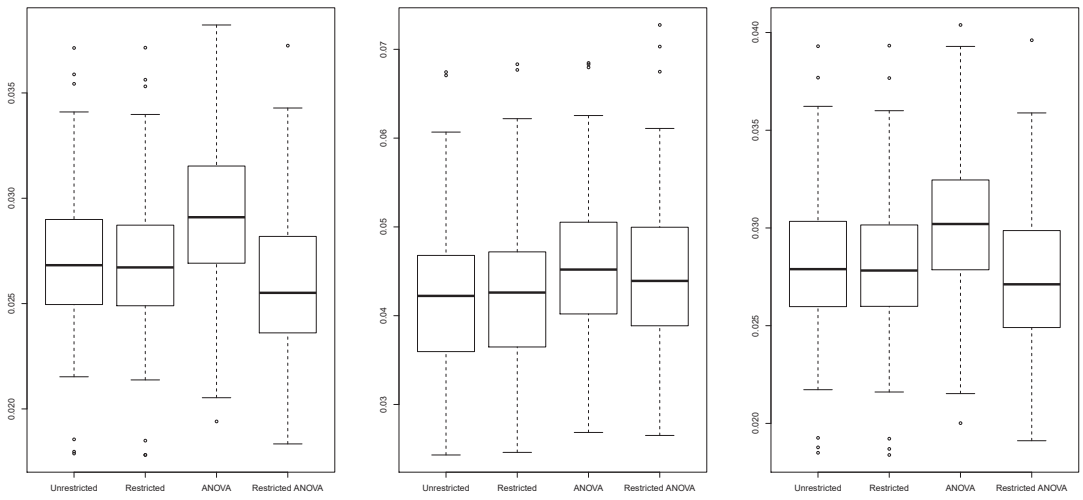


Figure C.1: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 1 and $n_{\boldsymbol{z}_p} = 0$ and $n_{\boldsymbol{x}_p} = 5$.

- $n_{\boldsymbol{z}_p} = 0$, $n_{\boldsymbol{x}_p} = 10$

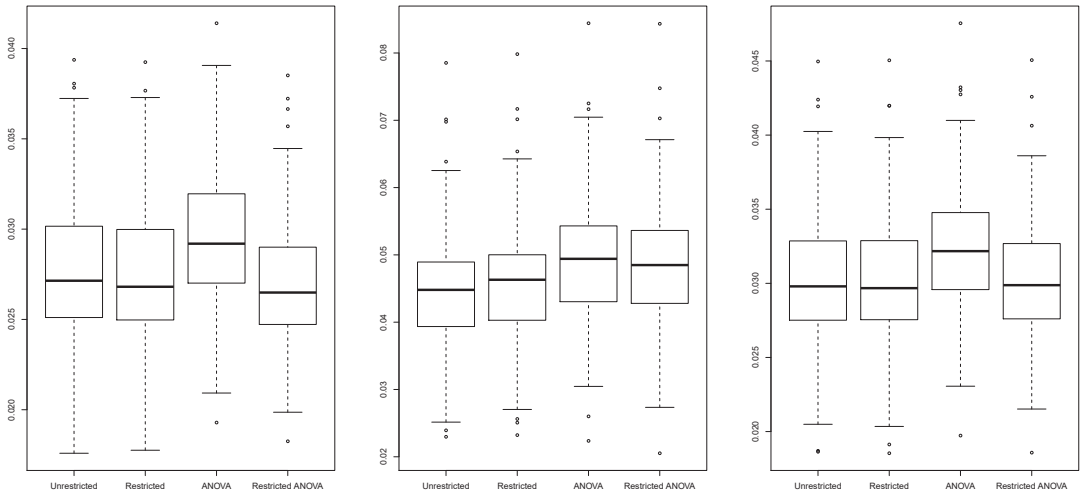


Figure C.2: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 1 and $n_{\boldsymbol{z}_p} = 0$ and $n_{\boldsymbol{x}_p} = 10$.

- $n_{\boldsymbol{z}_p} = 0$, $n_{\boldsymbol{x}_p} = 15$

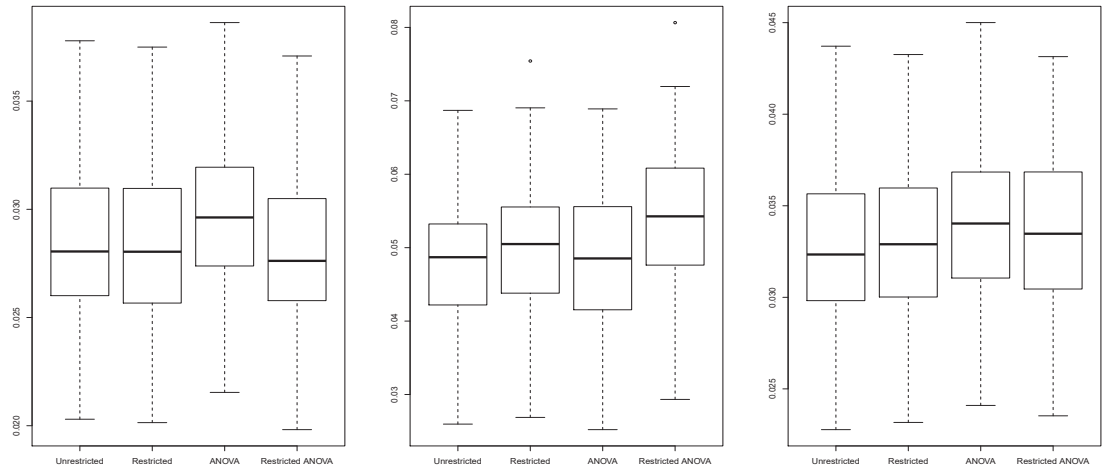


Figure C.3: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 1 and $n_{\boldsymbol{z}_p} = 0$ and $n_{\boldsymbol{x}_p} = 15$.

- $n_{\boldsymbol{z}_p} = 0$, $n_{\boldsymbol{x}_p} = 20$

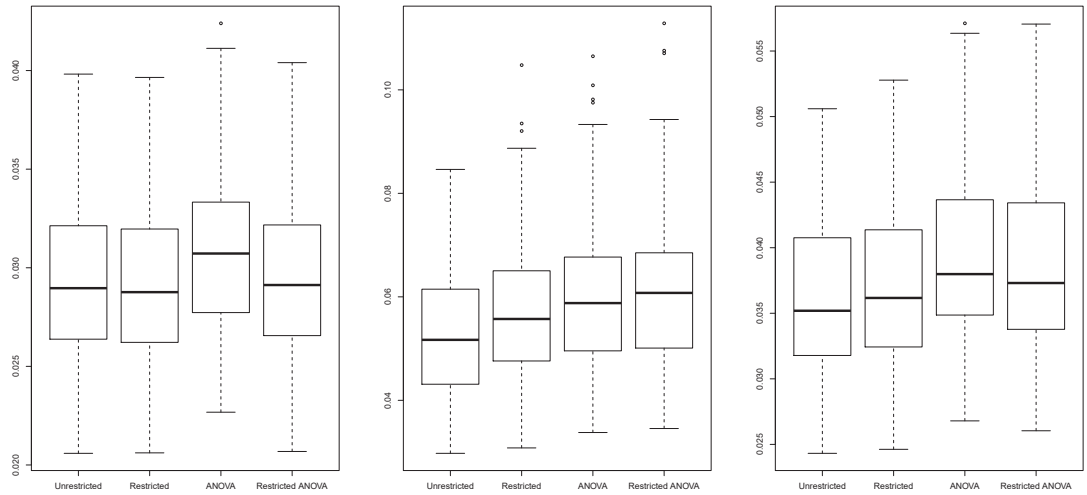


Figure C.4: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 1 and $n_{\boldsymbol{z}_p} = 0$ and $n_{\boldsymbol{x}_p} = 20$.

- $n_{\boldsymbol{z}_p} = 10$, $n_{\boldsymbol{x}_p} = 5$

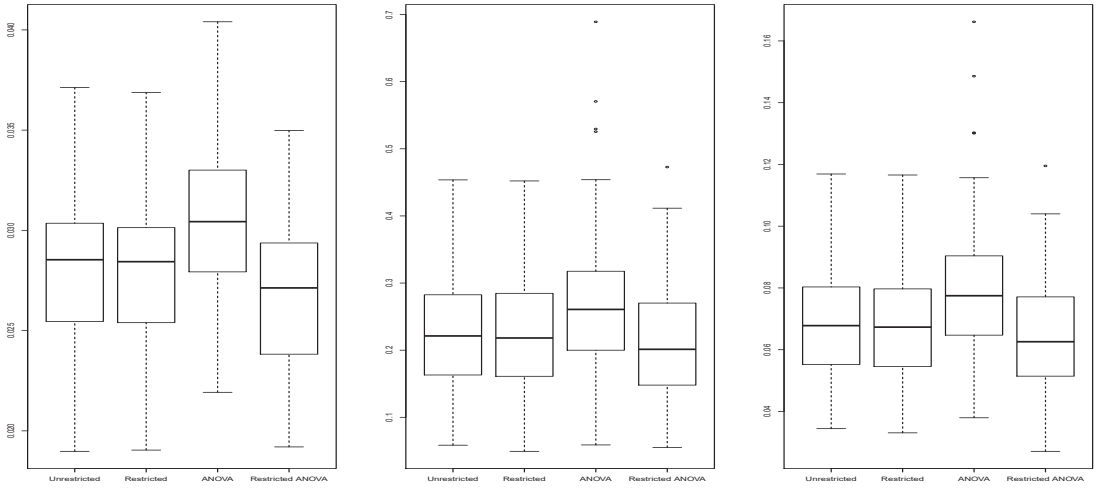


Figure C.5: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 1 and $n_{\boldsymbol{z}_p} = 10$, $n_{\boldsymbol{x}_p} = 5$.

- $n_{\boldsymbol{z}_p} = 10$, $n_{\boldsymbol{x}_p} = 15$

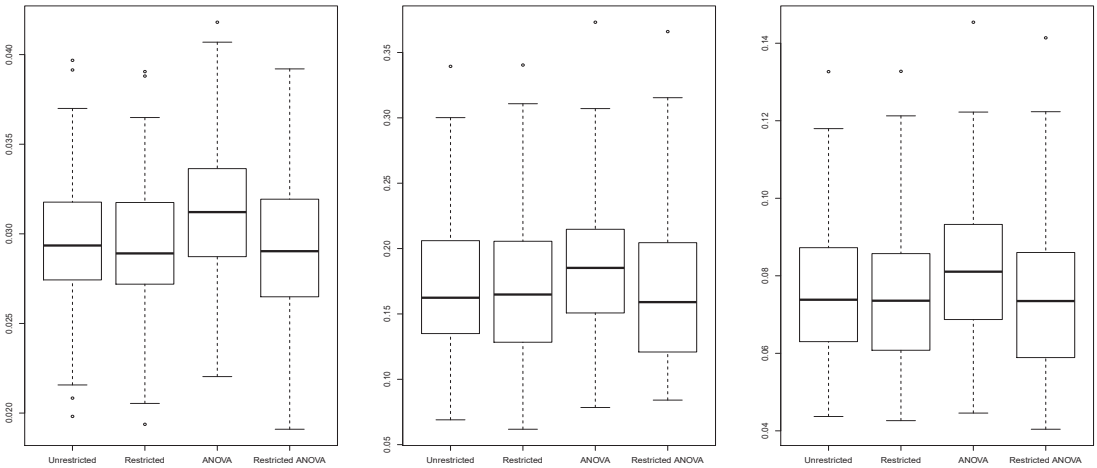


Figure C.6: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 1 and $n_{\boldsymbol{z}_p} = 10$, $n_{\boldsymbol{x}_p} = 15$.

- $n_{\boldsymbol{z}_p} = 10$, $n_{\boldsymbol{x}_p} = 20$

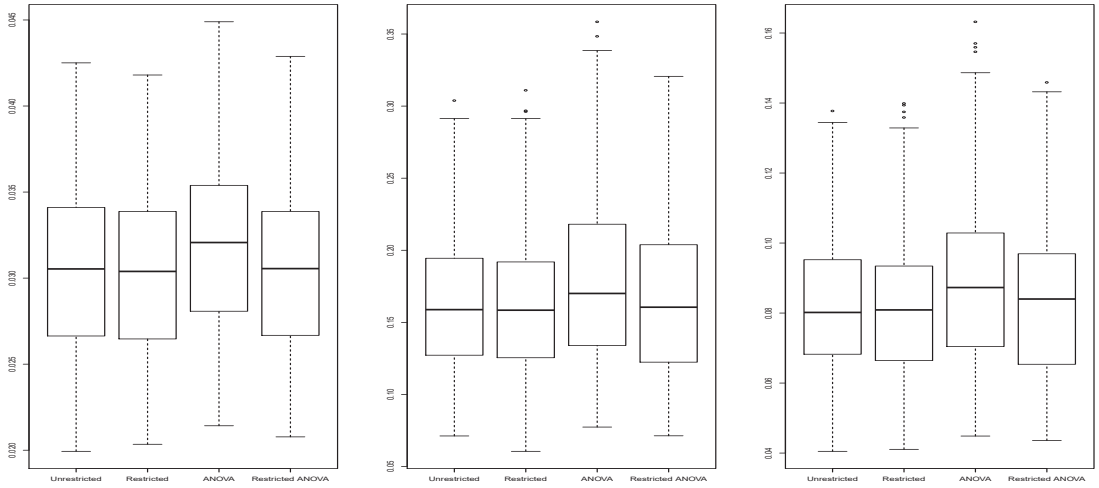


Figure C.7: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 1 and $n_{\boldsymbol{z}_p} = 10$, $n_{\boldsymbol{x}_p} = 30$.

- $n_{\boldsymbol{z}_p} = 20$, $n_{\boldsymbol{x}_p} = 5$

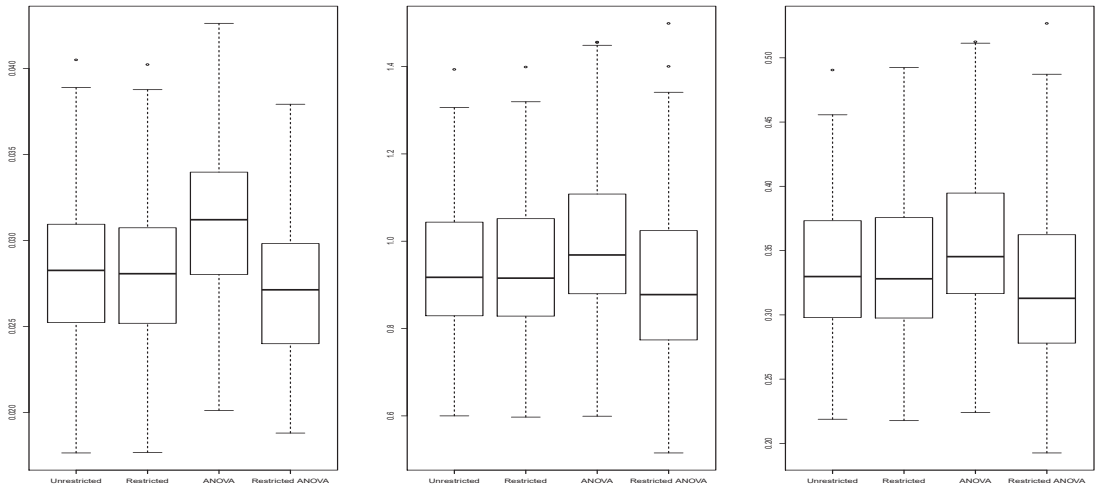


Figure C.8: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 1 and $n_{\boldsymbol{z}_p} = 20$, $n_{\boldsymbol{x}_p} = 5$.

- $n_{\boldsymbol{z}_p} = 20$, $n_{\boldsymbol{x}_p} = 10$

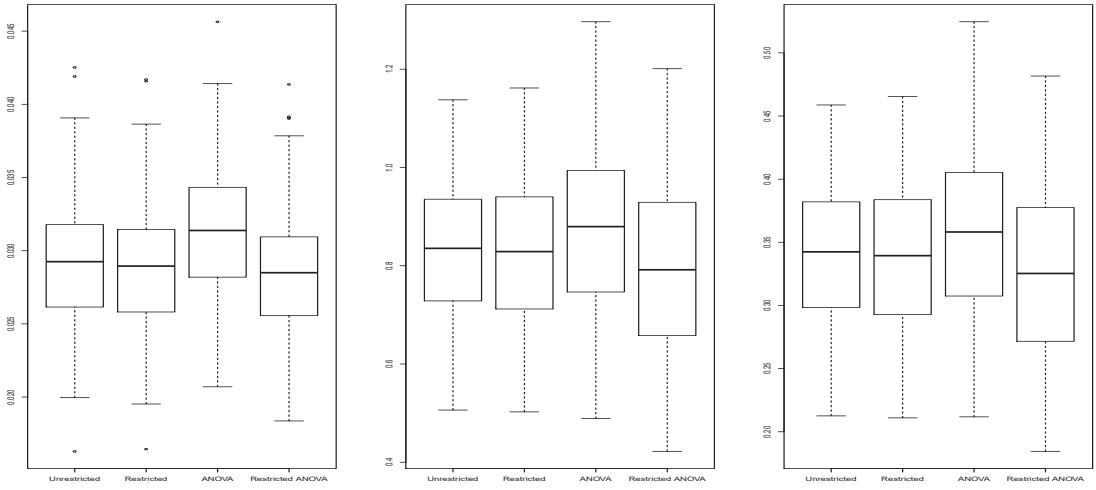


Figure C.9: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 1 and $n_{\boldsymbol{z}_p} = 20$, $n_{\boldsymbol{x}_p} = 10$.

- $n_{\boldsymbol{z}_p} = 20$, $n_{\boldsymbol{x}_p} = 15$

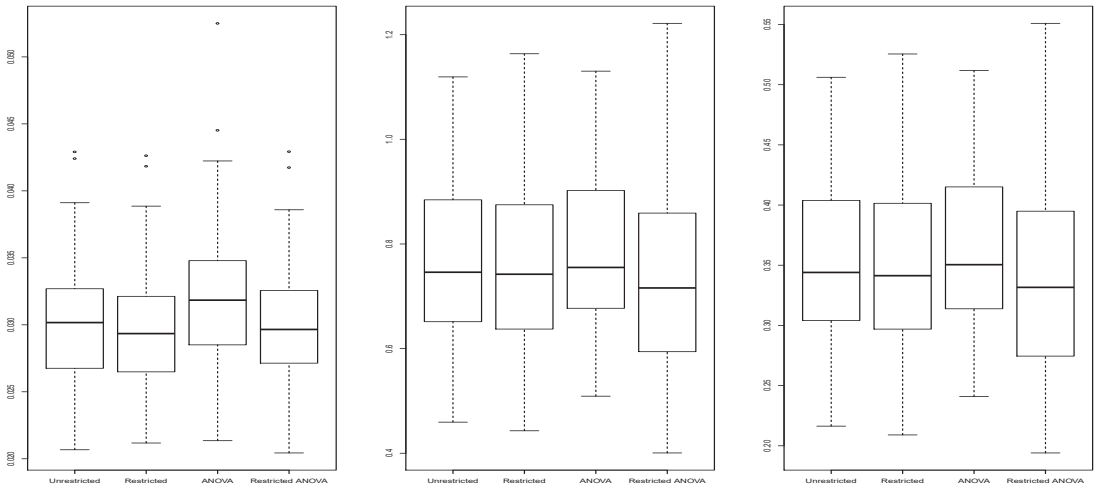


Figure C.10: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 1 and $n_{\boldsymbol{z}_p} = 20$, $n_{\boldsymbol{x}_p} = 15$.

- $n_{\boldsymbol{z}_p} = 20$, $n_{\boldsymbol{x}_p} = 20$

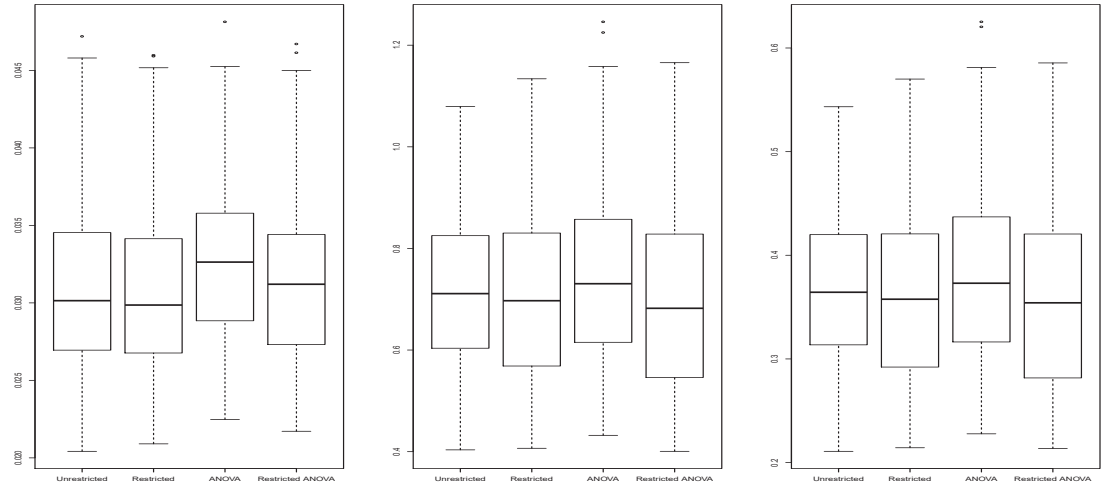


Figure C.11: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 1 and $n_{\boldsymbol{z}_p} = 20$, $n_{\boldsymbol{x}_p} = 20$.

Simulations results for Scenario 2:

- $n_{\boldsymbol{z}_p} = 0$, $n_{\boldsymbol{x}_p} = 5$

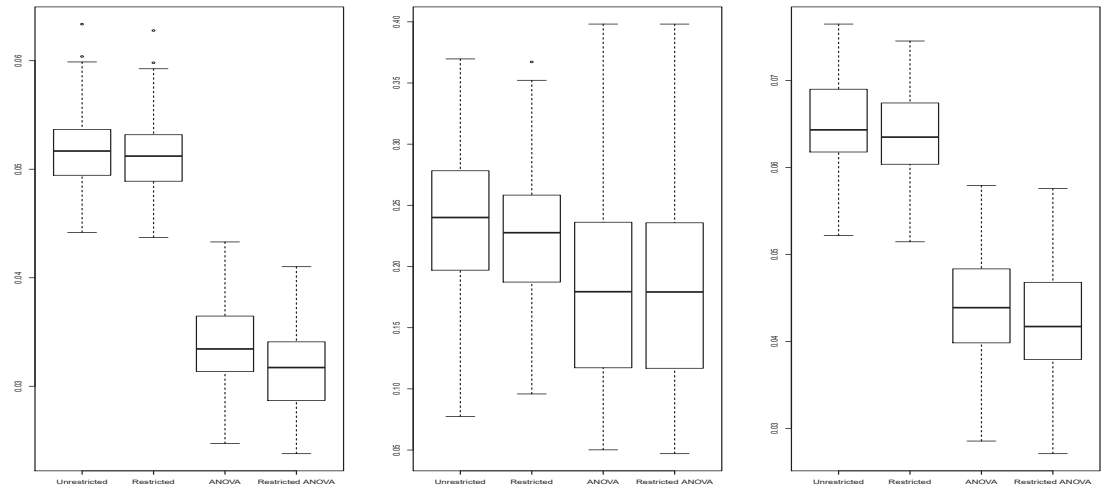


Figure C.12: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 2 and $n_{\boldsymbol{z}_p} = 0$ and $n_{\boldsymbol{x}_p} = 5$.

- $n_{\boldsymbol{z}_p} = 0$, $n_{\boldsymbol{x}_p} = 10$

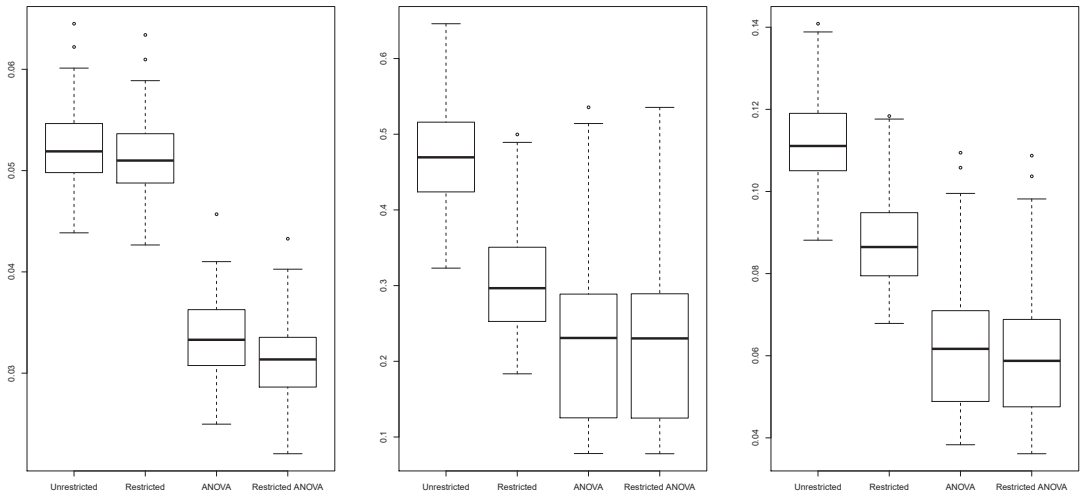


Figure C.13: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario and $n_{\boldsymbol{z}_p} = 0$ and $n_{\boldsymbol{x}_p} = 10$.

- $n_{\boldsymbol{z}_p} = 0$, $n_{\boldsymbol{x}_p} = 15$

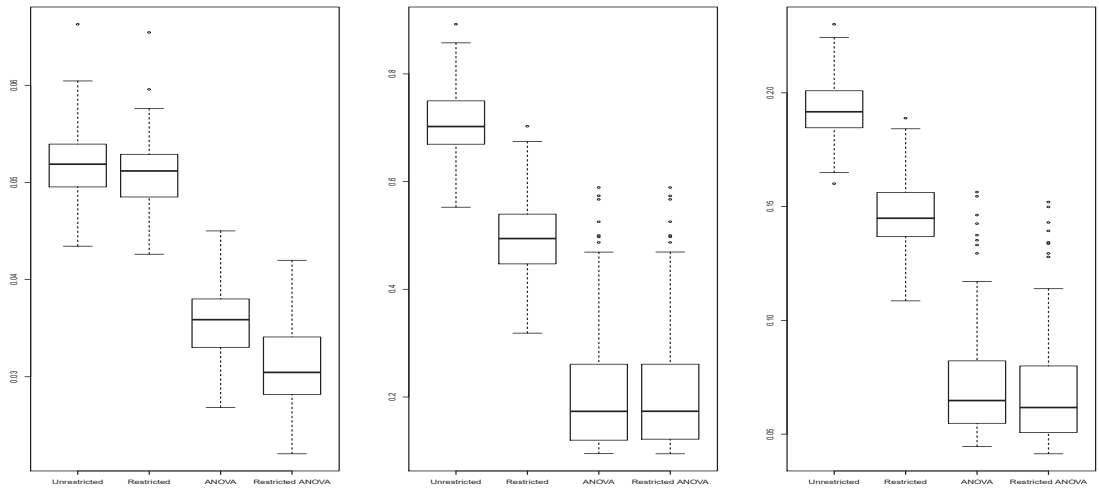


Figure C.14: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 2 and $n_{\boldsymbol{z}_p} = 0$ and $n_{\boldsymbol{x}_p} = 15$.

- $n_{\boldsymbol{z}_p} = 0$, $n_{\boldsymbol{x}_p} = 20$

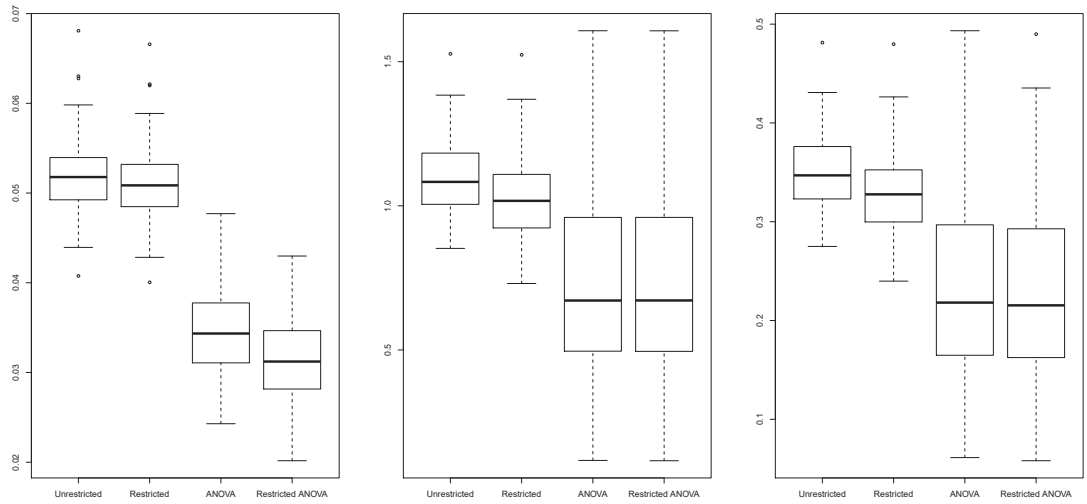


Figure C.15: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 2 and $n_{\boldsymbol{z}_p} = 0$ and $n_{\boldsymbol{x}_p} = 20$.

- $n_{\boldsymbol{z}_p} = 10$, $n_{\boldsymbol{x}_p} = 5$

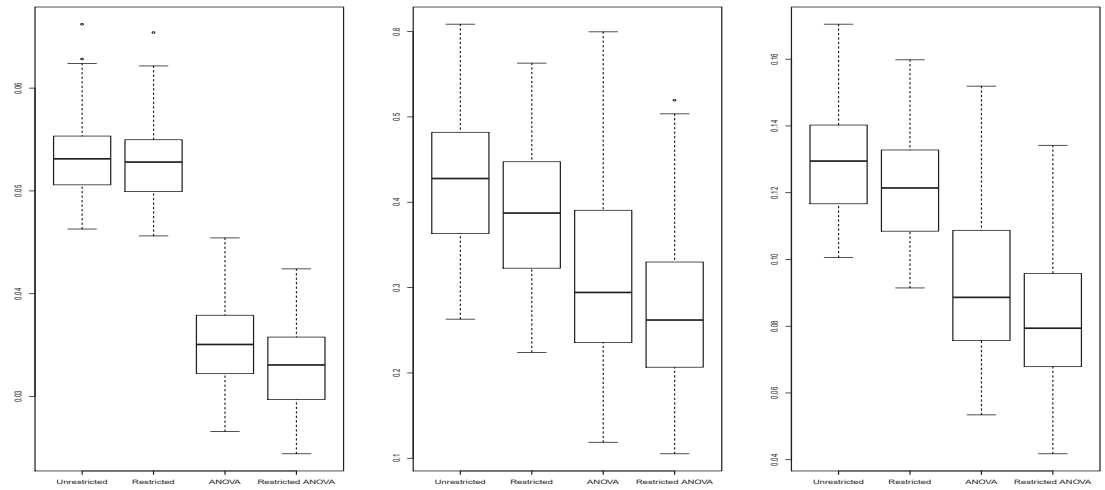


Figure C.16: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 2 and $n_{\boldsymbol{z}_p} = 10$ and $n_{\boldsymbol{x}_p} = 5$.

- $n_{\boldsymbol{z}_p} = n_{\boldsymbol{x}_p} = 10$

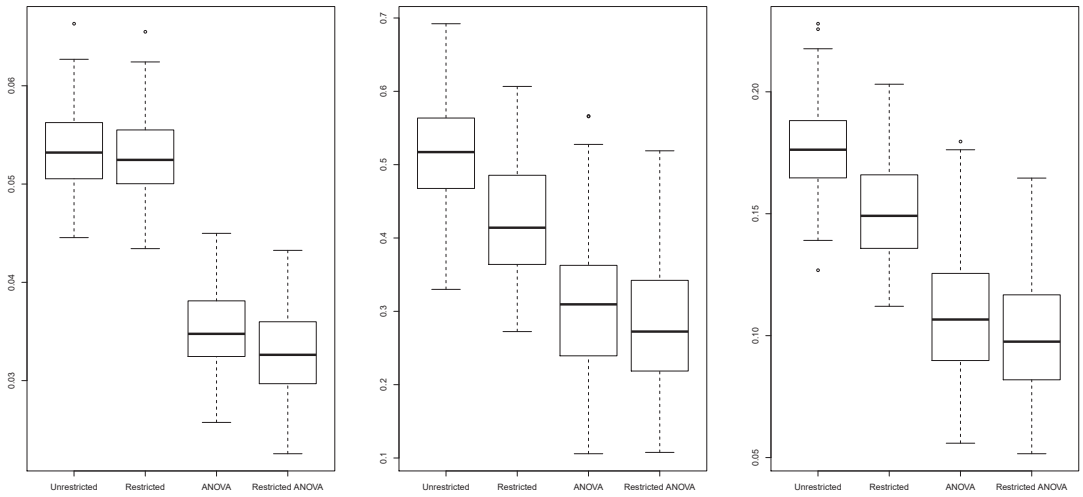


Figure C.17: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 2 and $n_{\boldsymbol{z}_p} = n_{\boldsymbol{x}_p} = 10$.

- $n_{\boldsymbol{z}_p} = 10$, $n_{\boldsymbol{x}_p} = 15$

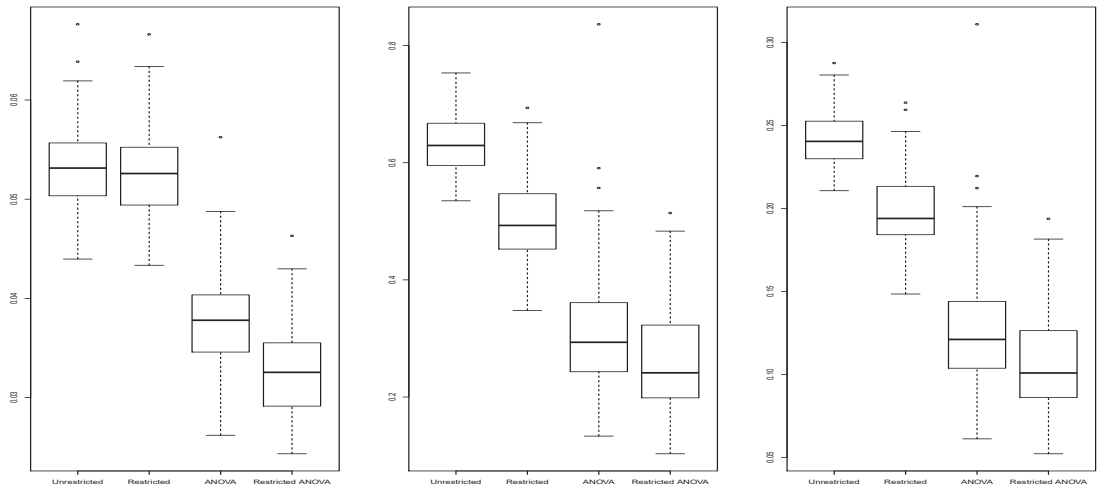


Figure C.18: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 2 and $n_{\boldsymbol{z}_p} = 10$ and $n_{\boldsymbol{x}_p} = 15$.

- $n_{\boldsymbol{z}_p} = 10$, $n_{\boldsymbol{x}_p} = 20$

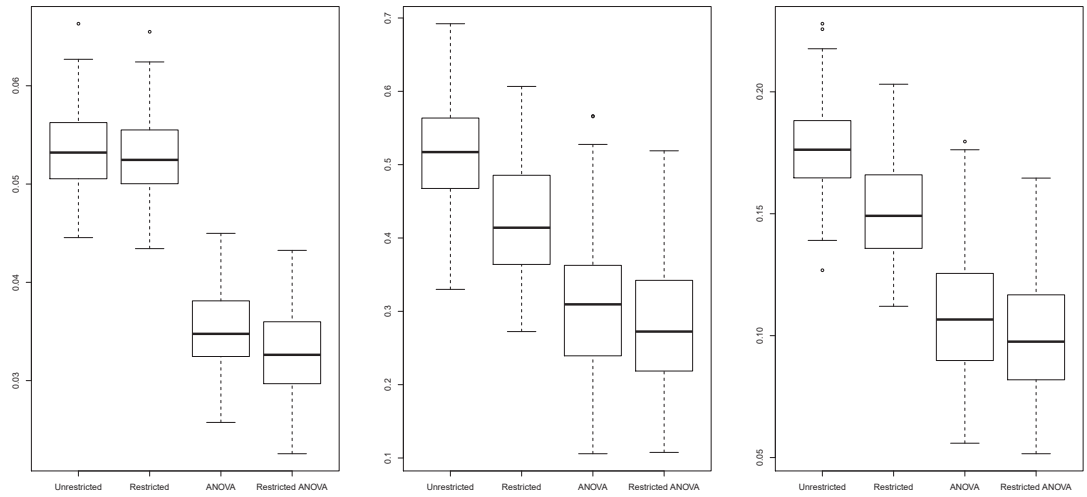


Figure C.19: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 2 and $n_{\boldsymbol{z}_p} = 10$ and $n_{\boldsymbol{x}_p} = 20$.

- $n_{\boldsymbol{z}_p} = 20$, $n_{\boldsymbol{x}_p} = 10$

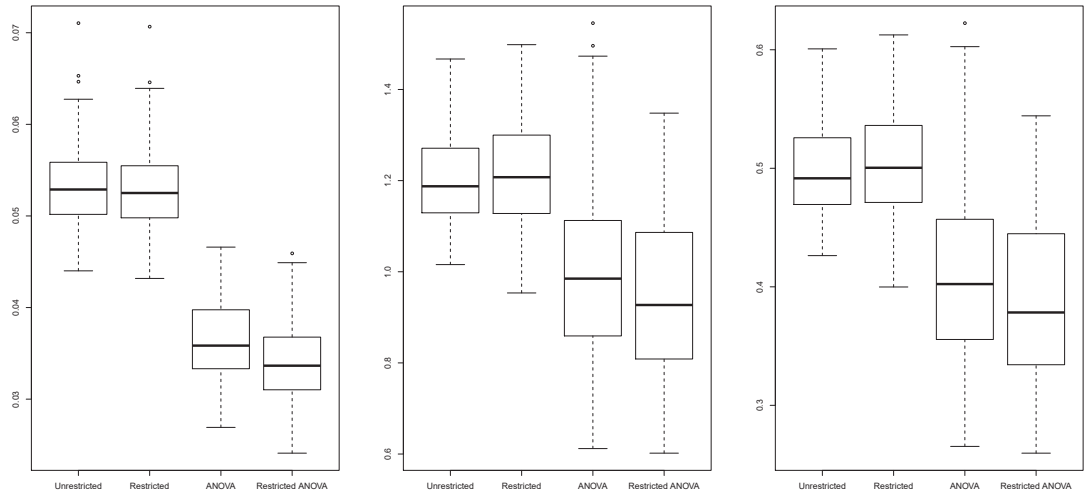


Figure C.20: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 2 and $n_{\boldsymbol{z}_p} = 20$ and $n_{\boldsymbol{x}_p} = 10$.

- $n_{\boldsymbol{z}_p} = 20$, $n_{\boldsymbol{x}_p} = 15$

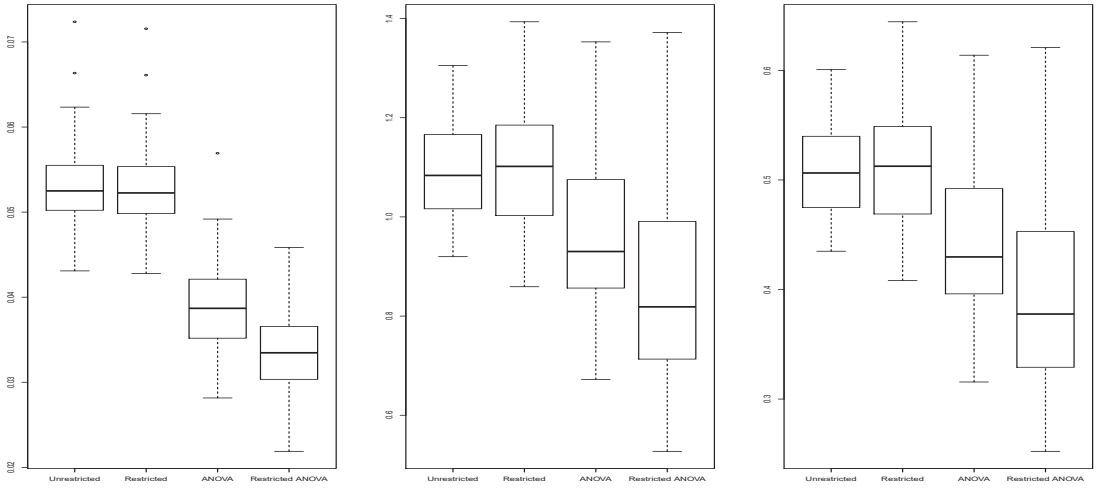


Figure C.21: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 2 and $n_{\boldsymbol{z}_p} = 20$ and $n_{\boldsymbol{x}_p} = 15$.

- $n_{\boldsymbol{z}_p} = n_{\boldsymbol{x}_p} = 20$

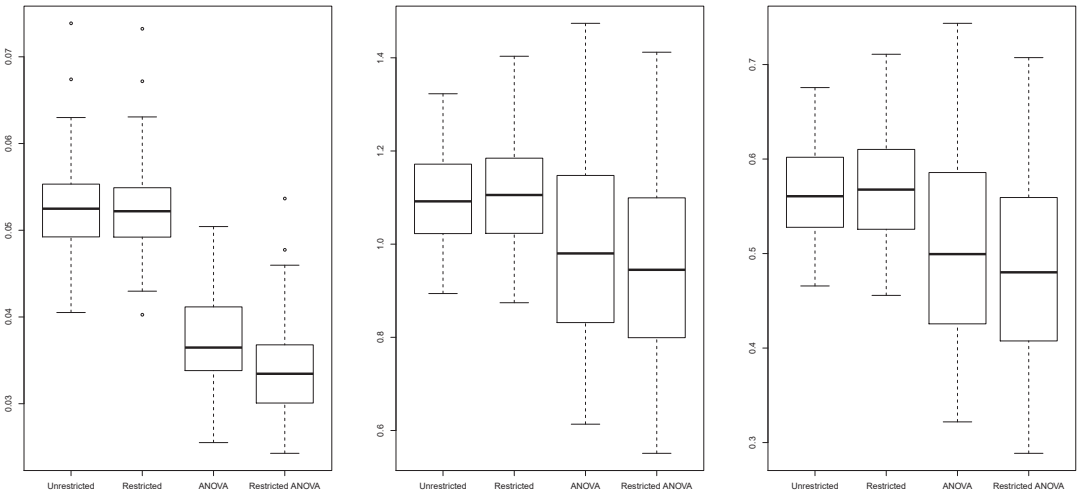


Figure C.22: MAE in the fit (left panel), in the out-of-sample prediction (middle panel) and in total (right panel) of smooth models in scenario 2 and $n_{\boldsymbol{z}_p} = n_{\boldsymbol{x}_p} = 20$.